ORIGIN AND EVOLUTION OF EUKARYOTIC GENE SEQUENCES DERIVED FROM TRANSPOSABLE ELEMENTS

A Dissertation
Presented to
The Academic Faculty

by

Jittima Piriyapongsa

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biology

Georgia Institute of Technology
August 2008

ORIGIN AND EVOLUTION OF EUKARYOTIC GENE SEQUENCES DERIVED FROM TRANSPOSABLE ELEMENTS

Approved by:

Dr. I. King Jordan, Advisor
School of Biology
Georgia Institute of Technology

Dr. John McDonald
School of Biology
Georgia Institute of Technology

Dr. Mark Borodovsky
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Leonid Bunimovich
School of Mathematics
Georgia Institute of Technology

Dr. Jung Choi
School of Biology
Georgia Institute of Technology

Date Approved: May 29, 2008

To my family...

ACKNOWLEDGEMENTS

First of all, I would like to take this opportunity to thank my advisor, Dr. I. King Jordan, for his constant guidance, motivation, and support. Without him, I would not have come this far. His knowledge, experience, and insights have been very influential in my research. Thank you so much for a great experience. Also, I would like to thank Dr. Mark Borodovsky, my first mentor here at Georgia Tech. I received much knowledge and experience from him as I worked in his lab for two years as a PhD student. The skills and scientific thought process I learned have been influential in my research. I would also like to thank my thesis committee members for their advice, time and expertise throughout this entire thesis process. Special thanks to my lab mates, Lee Katz and Ahsan Huda, for their friendships and for being supportive in every way. I also thank all Jordan's lab members for their help and support. Thanks to my fellow graduate students, Burcu Bakir and Navin Elango, for being my good friends since the first day we met. The five years here have been quite an experience and you all have made it a memorable time of my life.

I would like to especially thank my parents for their endless love, and support in all my efforts, and for giving me the foundation to be who I am. Without their encouragement and dedication, I would not be at this point. Even though my father is not here anymore, I always have him in my heart and I know that he would be proud of me. To my sisters - Yoo and Hui - thank you for always being there for me. Special thanks to my close friend, Nantachai Kantanantha (Lek), who encourages and supports me in every way since my first day as a PhD student. Thanks for always being here for me no matter what. I am also thankful to my Thai friends here - Gib, Nu, Oh, Chat, P'Golf, Jan, Chi,

Tam, Ter, P'Term for having dinner and sharing stories with me every Fridays. Also, I really enjoy every trips and every fun activites we had together here. Thank you to all my friends in Thailand for their support and understanding.

Finally, I would like to thank other people who deserve the gratitude but I have not mentioned here. To them, please accept my apology and thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

BLAST	Basic Local Alignment Search Tool
CCDS	Consensus CDS
cDNA	complementary DNA
CDS	Coding Sequence
CENP-B	Centromere protein B
CRF2	class II cytokine receptor family
CtBP	C-terminal Binding Protein
DAG	Directed Acyclic Graph
DBD	DNA-binding domain
dbEST	Expressed Sequence Tags Database
DNA	Deoxyribonucleic Acid
dsRNA	double stranded RNA
EMBL	The European Molecular Biology Laboratory
EN	Endonuclease
env	envelope
EST	Expressed Sequence Tag
GA	Gathering cutoff
gag	group-specific antigen
GO	Gene Ontology
HERV	Human Endogenous Retrovirus
HMM	Hidden Markov Model
IL22RA2	Interleukin 22 receptor, alpha 2
INT	Integrase

LINE	Long Interspersed Nuclear Element
LRTS	LTR RetroTransposon Sequence
LTR	Long Terminal Repeat
MaLR	Mammalian apparent LTR Retrotransposon
Mbp	Mega base pair
MEGA	Molecular Evolutionary Genetics Analysis
MIR	Mammalian-wide Interspersed Repeat
miRNA	microRNA
MITE	Miniature Inverted-repeat Transposable Element
MPSS	Massively Parallel Signature Sequencing
mRNA	messenger RNA
MYA	Million Years Ago
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
pol	polymerase
PTGS	Post-transcriptional Gene Silencing
RAG1	Recombination Activating Gene 1
RefSeq	Reference Sequence database
RISC	RNA Induced Silencing Complex
RNA	Ribonucleic Acid
RNAi	RNA interference
RT	Reverse Transcriptase

S/MARs	scaffold/matrix attachment regions
SAGE	Serial Analysis of Gene Expression
SETMAR	SET domain and Mariner transposase fusion gene
SINE	Short Interspersed Nuclear Element
siRNA	Small Interfering RNA
SW	Smith-Waterman
TC	Trusted cutoff
TE	Transposable Element
TIR	Terminal Inverted Repeat
UBP1b	Oligouridylate binding protein1b gene
UCSC	University of California Santa Cruz
UTR	Untranslated Region
VLP	Virus-like particle

SUMMARY

Transposable elements (TEs), or mobile genetic elements, are major components of eukaryotic genomes. TEs became of great interest over the last few decades because of their significant impact on gene and genome evolution. My dissertation encompasses five different studies that are linked by a common theme – the investigation of TE contributions to eukaryotic gene sequences. The studies focus on two types of gene sequences: protein coding genes and non-coding regulatory genes. The instances, causes, and consequences of TE integration into human protein coding genes have been studied previously. However, the precise extent of TE contribution to host protein coding sequences and the coding potential of such TE-derived sequences remain a matter of controversy. The first objective of this dissertation is to investigate the extent, evolution, and coding property of TE-derived protein-coding sequences in human genes as well as the ascertainment bias of methods used to detect such sequences.

Small noncoding regulatory RNAs, such as microRNAs (miRNAs) and short interfering RNAs (siRNAs), are a class of genes which was recently discovered. Accordingly, a number of open questions regarding their evolutionary origins remain. siRNAs are known to originate from and regulate TEs, whereas miRNAs are encoded from distinct genetic loci and thought to be dedicated to the regulation of host genes. The second objective of my dissertation research objective is to explore the extent of TE contributions to the origin and evolution of miRNA genes including the possible evolutionary connection between the origin of both siRNAs and miRNAs.

The results from my research provide the following five major advances to the study of TE-gene evolution:

Research advance1: The first detailed analysis of exonization events of one particular class of TE, long terminal repeat (LTR) containing elements, in the human genome indicates that 5.8% of human genes are associated with LTR elements and 50 distinct protein coding exons were comprised exclusively of LTR retrotransposon sequences. A detailed scenario of the exonization process of an alternatively spliced exon of the alpha 2 gene of the Interleukin 22 receptor (*IL22RA2*) was supported by new experimental data generated in this research. As a result of a single mutation in the proto-splice site, recruitment of the part of MaLR LTR as a novel exon in great ape species occurred prior to the divergence of orangutans and humans from a common ancestor (~ 14 MYA). The majority of human LTR exonization events involve non-coding exon sequences in the 5' and 3' untranslated regions.

Research advance2: Differences in the extent of TEs found in experimentally characterized protein sequences (CDS) caused by the specific bias of each search method are emphasized by the comparison of the results from three sequence similarity search approaches: 1-a profile-based approach, 2-BLAST methods and 3-RepeatMasker. Profile based methods show a valuable combination of sensitivity, measured by their ability to detect well-characterized cases of TE-derived CDS, and selectivity compared to the other methods evaluated. The non-overlap of hits and difference in the ability of each approach to recover known cases of TE-derived CDS indicates the need to use these complementary methods together for more accurate detection of CDS that evolved from TEs. On average, TE-derived exon sequences have low protein coding potential. In

particular, non-coding TEs, such as Alu elements, are frequently exonized but unlikely to encode protein sequences. I hypothesize that many of these non-coding exonized TEs are involved in gene regulation via the formation of double stranded RNA (dsRNA) complexes with complementary TE-derived exons.

Research advance3: The investigation of the relationship between human miRNAs and TEs shows that 55 experimentally verified human miRNA genes (~12%) originated from TEs. Sequence comparisons among vertebrate genomes revealed that, on average, TE-derived human miRNAs are significantly less conserved than non TE-derived miRNAs. However, there are TE-derived miRNAs that are relatively conserved, and most are related to the ancient L2 and MIR families. In addition to the known human miRNAs that were shown to be derived from TE sequences, an additional 85 novel TE-derived miRNA genes were predicted in this study. The dispersed repetitive nature of TE sequences provides for the emergence of multiple novel miRNA genes as well as numerous homologous target sites throughout the genome. Thus, TEs may represent a mechanism for the rapid deployment of miRNA based regulatory networks in the human genome.

Research advance4: A group of seven closely related miRNA genes (hsa-mir-548) was found to be derived from the Made1 family of MITEs. These Made1 elements are nearly perfect palindromes which are able to form highly stable hairpin-loops, resembling pre-miRNA structures. The analysis of their expression profiles and functional affinities suggests cancer-related regulatory roles for hsa-mir-548.

Research advance5: An original model for a siRNA-to-miRNA evolutionary transition mediated by DNA-type TEs is proposed. This model is supported by the

presence of evolutionary intermediate TE sequences that encode both siRNAs and miRNAs in the Arabidopsis and rice genomes. These dual coding TEs can be expressed as read-through transcripts from the intronic regions of spliced RNA messages. The results indicate that ancestral miRNAs could have evolved from TEs prior to the full elaboration of the miRNA biogenesis pathway. The siRNA-to-miRNA evolutionary transition is representative of a number of other regulatory mechanisms that evolved to silence TEs and were later co-opted to serve as regulators of host gene expression.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Overview of Transposable Elements (TEs)

Transposable elements (TEs) are mobile DNA sequences that have ability to move (transpose) from one location to another within genomes of their host organisms and often make duplicate copies of themselves in the process. Since their first discovery by Barbara McClintock about 60 years ago (MCCLINTOCK 1948), TEs have been found to be the largest component of the genetic material of most eukaryotes. Eukayotic species vary considerably in the distribution and the content of their TE sequences which can account for the main differences in genome size among species (BIEMONT and VIEIRA 2005; BIEMONT and VIEIRA 2006; KIDWELL 2002) (Table 1.1). Of all eukaryotic genomes sequenced to date, only *Plasmodium falciparum* genome was found not to host any active TEs (GARDNER *et al.* 2002).

TEs can be classified into two major classes based on modes of transposition (FINNEGAN 1989; FINNEGAN 1992). Class I elements (retrotransposons) transpose through RNA intermediates. They are transcribed into RNA, and then reverse transcribed back to DNA and reintegrated into the genome, thereby duplicating the element ("copy and paste" mechanism). Reverse transcription and integration are catalyzed by reverse transcriptase (RT) and endonuclease/integrase (EN/INT), which are encoded by autonomous elements. Retrotransposons can be further divided into two subclasses based on the presence/absence of long terminal repeats (LTRs).

Table 1.1: Genome sizes and transposable element proportions for eukaryotic species

Scientific name	Common name	Genome size (Mbp)	%TE	References
Lilium	Lilies	36,000	95-99	(BENNETZEN 2000; FLAVELL *et al.* 1994; FLAVELL *et al.* 1974)
Zea Mays	Maize	2,500	60-80	(SANMIGUEL *et al.* 1996)
Hordeum vulgare	Barley	4,800	55	(KUMAR and BENNETZEN 1999; VICIENT *et al.* 1999)
Macaca mulatta	Rhesus monkey	3,100	50	(HAN *et al.* 2007)
Homo sapiens	Human	3,200	45	(LANDER *et al.* 2001)
Pan troglodytes	Chimpanzee	3,000	45	(MIKKELSEN *et al.* 2005)
Bos taurus	Cow	3,200	40	(LARKIN *et al.* 2003)
Rattus norvegicus	Rat	2,500	40	(GIBBS *et al.* 2004)
Mus musculus	Mouse	3,000	38	(WATERSTON *et al.* 2002)
Xenopus laevis	African clawed frogs	3,100	37	(CARROLL *et al.* 1989)
Canis Familiaris	Dog	2,400	34	(LINDBLAD-TOH *et al.* 2005)
Oryza sativa	Rice	430	20	(TURCOTTE *et al.* 2001)
Anopheles gambiae	Mosquito	278	16	(HOLT *et al.* 2002)
Drosophila melanogaster	Fruit fly	180	15-22	(ADAMS *et al.* 2000; VIEIRA *et al.* 1999)
Arabidopsis thaliana	Thale cress	130	14	(THEARABIDOPSISGENOMEINITIATIVE 2000)
Dictyostelium discoideum	Slime mold	34	10	(GLOCKNER *et al.* 2001)
Gallus gallus	Chicken	1,050	9	(INTERNATIONALCHICKENGENOMESEQUENCINGCONSORTIUM 2004)
Caenorhabditis elegans	Worm	103	6	(THEC.ELEGANSSEQUENCINGCONSORTIUM 1998)

Table 1.1 continued

Scientific name	Common name	Genome size (Mbp)	%TE	References
Drosophila simulans	Vinegar fly	142	5	(VIEIRA *et al.* 1999)
Saccharomyces cerevisiae	Baker's yeast	12	3-5	(KIM *et al.* 1998)
Fugu rubripes	Japanese pufferfish	393	2.7	(APARICIO *et al.* 2002)
Schizosaccharomyces pombe	Fission yeast	12.2	1.1	(BOWEN *et al.* 2003)
Tetraodon nigroviridis	Green spotted puffer	342	0.14	(CROLLIUS *et al.* 2000; DASILVA *et al.* 2002)

LTR retrotransposons contain direct repeats at both ends and are related, in sequence and genomic structure, to retroviruses; they both contain *gag* and *pol* genes that encode all necessary proteins to provide enzymatic activities for transposition process (Figure 1.1). They differ in that retroviruses encode an envelope protein, whereas LTR retrotransposons either lack or contain a remnant of an *env* gene.

Non-LTR retrotransposons lack LTRs and possess a polyadenylate sequences at their 3' termini. They are divided into two super-families, LINEs (Long Interspersed Nuclear Elements) and SINEs (Short Interspersed Nuclear Elements). LINEs are autonomous retroelements encoding two ORFs, ORF1 encoding an RNA binding protein and ORF2 encoding endonuclease and reverse transcriptase activities (Figure 1.1). SINEs are non-autonomous retroelements which lack coding capacity and are dependent on the reverse transcriptase machinery encoded by LINEs for their mobility (DEWANNIEUX *et al.* 2003).

Class II elements (DNA transposons) transpose directly as DNA sequence. These elements are generally excised from one genomic site and integrated into another by a

conservative "cut and paste" mechanism catalyzed by the enzyme transposase. They have terminal inverted repeats (TIRs) flanking an ORF encoding a transposase (Figure 1.1). Non-autonomous DNA elements containing only TIRs may be transposed in *trans* by related full length autonomous elements.

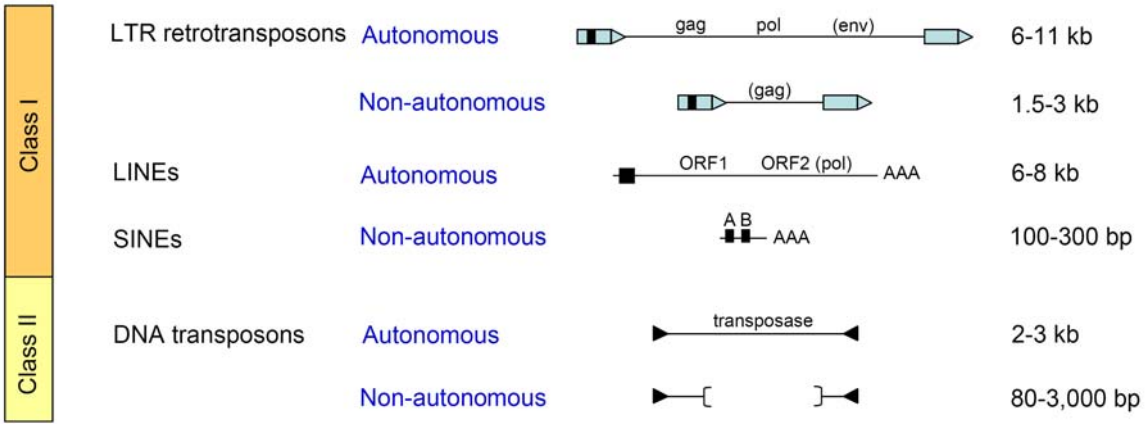

Figure 1.1: Classes of TEs found in human genome and their characteristics

Transposable elements: "junk DNA" or "genetic treasure"

Eukaryotic genomes are very complex and dynamic systems. Only a small fraction of eukaryotic genomes consists of protein coding exons, while a far more substantial part is constituted by TEs. For human, TEs contribute more than 45% of the whole genome (JASINSKA and KRZYZOSIAK 2004; LANDER *et al.* 2001). An interesting question is whether the abundance and persistence of TEs in eukaryotic genomes rest primarily on their ability to out-replicate their host genome or on the contributions they make to the evolution and genetic plasticity of their hosts (CHARLESWORTH *et al.* 1994; MCDONALD 1995). TEs have long been considered as "junk" or "selfish" DNA parasitizing the genome of living organisms (DOOLITTLE and SAPIENZA 1980; HICKEY

1982; OHNO 1972; ORGEL and CRICK 1980). The selfish DNA theory (DOOLITTLE and SAPIENZA 1980; ORGEL and CRICK 1980) describes that TEs serve only to survive and increase their number even at the expense of their host genomes. They are maintained in the host genomes by natural selection for their ability to propagate themselves without providing any function or benefit to their hosts. According to the selfish DNA theory, TEs make no positive contribution to host function, phenotype or evolution, and the effects that TEs have on their hosts are harmful and selected against. For instance, TEs can insert into host genes or gene regulatory elements and this will lead to deleterious effects including hereditary diseases induced by insertion mutations. The ability to cause such insertional mutations has been cited as a possible role for TEs in certain types of cancer (SAUTER *et al.* 1995; WANG-JOHANNING *et al.* 2003). Approximately 0.5–1% of human illnesses were believed to be associated with gene dysfunction or misregulation by TEs (KAZAZIAN 1998). TEs can also affect stability of the genome by introducing recombination hot spots (MIGHELL *et al.* 1997; SCHWARTZ *et al.* 1998).

Although TEs are considered to be harmful and are expected to be removed by the force of natural selection, they cannot be easily eliminated and their endurance in the host may be attributable to some evolutionary advantages they provide. The idea that TEs are beneficial to the host was proposed before (BROSIUS 1991; HARTL *et al.* 1983; KIDWELL and LISCH 2001; MCCLINTOCK 1984). As a source of genomic variation, TEs have potential to play a major role in the evolution of host genomes. TEs frequently donate regulatory and coding sequences to host genes (JORDAN *et al.* 2003; NEKRUTENKO and LI 2001; VAN DE LAGEMAAT *et al.* 2003; VOLFF 2006). TEs are also considered as the driving force in the evolution of epigenetic regulation (LIPPMAN *et al.* 2004) and in

speciation (ROSE and DOOLITTLE 1983). The preservation of TEs in the host genome is enhanced by some mechanisms that TE have evolved to alleviate the harmful effects caused by transposition. For example, the restriction of expression of some TEs to germline tissue (TRELOGAN and MARTIN 1995) decreases the detrimental effects of somatic mutations. Another example is a repressor protein of P-elements in Drosophila that blocks transposition in already infected genomes (ROBERTSON and ENGELS 1989).

More and more evidence regarding the role of TEs in host genomes has accumulated in recent years and enhanced our understanding about the impact of these elements on genome evolution (BIEMONT and VIEIRA 2006; JURKA *et al.* 2007; KAZAZIAN 2004; KIDWELL and LISCH 2001). However, an ongoing debate still exists regarding whether TEs actually confer a benefit to the host organism. A precise delineation of underlying mechanisms responsible for TE-host co-evolution remains elusive and needs to be clarified.

Contribution of transposable elements to host protein coding sequences

One of the many ways that TEs participate in the function and evolution of the resident genomes is through the donation of host protein coding sequences (CDSs). A specific kind of exaptation termed 'molecular domestication' has been observed for both major classes of TEs, the retrotransposons and the DNA transposons. Exaptation is a general term that describes an evolutionary event whereby any organismic feature takes on a functional role that is distinct from the function to which it was originally adapted (GOULD and VRBA 1982). Molecular domestication is defined as the process whereby a formerly selfish or parasitic TE is co-opted (exapted) to perform a function that benefits its host genome (MILLER *et al.* 1992). Approximately 50–100 protein-coding genes in

mammalian genomes evolved from coding sequences of TEs through these processes (BRANDT *et al.* 2005b; CAMPILLOS *et al.* 2006; KAPITONOV and JURKA 2004; KAPITONOV and JURKA 2005; KAPITONOV *et al.* 2004; LANDER *et al.* 2001; VOLFF 2006) (Table 1.2). Interestingly, most of these genes were derived from TE transposases.

There are several well known cases of TE-derived CDS. The RAG1, which catalyzes the V(D)J recombination necessary for the assembly of immunoglobulin and T-cell-receptor genes in developing lymphocytes (SCHATZ *et al.* 1989; TONEGAWA 1983), is probably the most ancient known host protein derived from a transposase of DNA-type elements (KAPITONOV and JURKA 2004; KAPITONOV and JURKA 2005). The telomerase enzyme, a reverse transcriptase involved in the replication of telomeres had its origin from the reverse transcriptase of a retrotransposable element (EICKBUSH 1997; NAKAMURA *et al.* 1997). Other well-characterized examples include the centromere binding protein CENP-B, which is related to pogo-like DNA transposases (KIPLING and WARBURTON 1997).

TEs insert into an ORF by two possible mechanisms: directly by transposition into an exon or indirectly by recruiting an intronic TE. It appears that the latter scenario of splice-mediated insertion of an intronic element is the major mechanism by which TEs are introduced into CDS (NEKRUTENKO and LI 2001). The underlying mechanisms of the TE insertion into the ORF of a host gene were discussed in detail previously (MAKALOWSKI 1995; MAKALOWSKI *et al.* 1994). The exonization process is facilitated by sequence motifs that resemble splice sites (pseudo splice sites), mostly found in the minus strand of Alu elements (MAKALOWSKI 2000; MAKALOWSKI *et al.* 1994; SOREK *et al.* 2002).

Although a number of TE sequences that have been domesticated to provide host CDSs have been identified, the actual proportion of host CDSs that were derived from TEs is still a matter of controversy. A number of large-scale analyses have been made in an attempt to exhaustively characterize instances of TE-derived host CDS (BRITTEN 2006; JURKA and KAPITONOV 1999; LANDER *et al.* 2001; NEKRUTENKO and LI 2001; SMIT 1999). Before the availability of complete human genome sequence, 20 cases of human CDSs derived from TEs were identified (JURKA and KAPITONOV 1999; SMIT 1999). Later, investigation of the draft sequence of the human genome found 47 cases including previously discovered ones, accounting for a total of 0.16% of human genes (LANDER *et al.* 2001). In the same year, Nekrutenko and Li, analyzed about 14,000 protein coding genes using similar techniques and reported that ~4% of analyzed human genes had a TE sequence present within a CDS (NEKRUTENKO and LI 2001). However, no evidence of TE-derived sequences was found when the same detection program was applied to a more reliable data set with supported evidence of protein expression and function based on 3D structures (PAVLICEK *et al.* 2002). Although a more sensitive technique did found some cases of TE-derived CDS, but none of these were from Alu elements, which lack protein coding capacity. Indeed, the actual protein coding potential of these non-coding TE sequences is still a matter of speculation. Most of well-supported cases of TE-derived host CDSs come from TEs that already encode proteins. Recently, protein sequence comparisons have been shown to be more sensitive than DNA comparison used in previous studies for detection of TE-derived host CDSs. For instance, protein sequence similarity searches were shown to detect more than twice as many cases of TE-derived CDS than DNA sequence similarity based search techniques in

one recent study (BRITTEN 2006). In general, the contradictory results obtained from these previous studies emphasize the differences in sensitivity and specificity for each of the search methods used. In any case, an exhaustive elucidation of the extent and significance of TE-derived host CDSs will be important for better understanding of the evolutionary dynamic between TEs and their host genomes.

Table 1.2: Protein-coding host genes domesticated from TEs [information from (VOLFF 2006)]

Gene	Protein functions	TE gene	Organisms	References
Syncytin-1	membrane glycoprotein, cell fusion, placenta formation	envelope	human	(MI *et al.* 2000)
Syncytin-2	membrane glycoprotein, cell fusion, placenta formation	envelope	human	(BLAISE *et al.* 2003)
Syncytin-A/B	membrane glycoprotein, cell fusion, placenta formation	envelope	mouse	(DUPRESSOIR *et al.* 2005)
Other *env* genes	unknown	envelope	mammals	(BLAISE *et al.* 2005)
Iris	defence against viruses?	envelope	*Drosophila*	(MALIK and HENIKOFF 2005)
Iris-like	defence against viruses?	envelope	mosquitoes	(MALIK and HENIKOFF 2005)
Peg10/ Mart2	cell proliferation, transcription factor?, placenta formation	gag	mammals	(ONO *et al.* 2006)
Ldoc1/ Mart7	inhibition of NF-kappaB activation, induction of apoptosis	gag	mammals	(NAGASAKI *et al.* 2003)
Other *Mart* genes	unknown	gag	mammals	(BRANDT *et al.* 2005a; BRANDT *et al.* 2005b; POULTER and BUTLER 1998; SEITZ *et al.* 2003; YOUNGSON *et al.* 2005)
Map-1/ Ma1	Bax-associating protein, induction of apoptosis	gag	mammals	(DALMAU *et al.* 1999; TAN *et al.* 2001)
Other *Ma* genes	neuronal autoantigens in paraneoplastic neurological diseases	gag	mammals	(SCHULLER *et al.* 2005; WILLS *et al.* 2006)

Table 1.2 continued

Gene	Protein functions	TE gene	Organisms	References
Fv1	murine leukemia virus restriction	gag	mouse	(BEST *et al.* 1996)
Gin-1	unknown	integrase	mammals	(LLORENS and MARIN 2001)
c-integrase	unknown	integrase/ transposase	mammals	(FESCHOTTE and PRITHAM 2005; GAO and VOYTAS 2005)
Fob1p	replication terminator protein, rDNA metabolism	integrase/ transposase	baker's yeast	(DLAKIC 2002; FESCHOTTE and PRITHAM 2005; GAO and VOYTAS 2005)
Telomerase	ribonucleo- protein, reverse transcriptase, telomere replication	reverse transcriptase	eukaryotes	(LINGNER *et al.* 1997; NAKAMURA *et al.* 1997)
Rag1	nuclease, V(D)J recombination, adaptive immune system	transposase	vertebrates	(AGRAWAL *et al.* 1998; HIOM *et al.* 1998; KAPITONOV and JURKA 2005; ZHOU *et al.* 2004)
Daysleeper	DNA binding protein, gene regulation, plant development	transposase	*Arabidopsis*	(BUNDOCK and HOOYKAAS 2005)
Gary	unknown	transposase	cereal grasses	(MUEHLBAUER *et al.* 2006)
Tram	unknown	transposase	mammals	(ESPOSITO *et al.* 1999)
Zbed4	unknown	transposase	mammals	(SMIT 1999)
KIAA0543	unknown	transposase	mammals	(SMIT 1999)
P52rIPK	binding to inhibitor of interferon-induced protein kinase PKR	transposase	mammals	(GALE *et al.* 1998)
Buster1-3	unknown	transposase	mammals	(SMIT 1999)
Dref	transcription factor, regulation of DNA replication	transposase	*Drosophila*	(ROBERTSON 2002)
Lin-15B	regulator of vulval development	transposase	*Caenorhabditis*	(ROBERTSON 2002)
Cenp-b	DNA binding protein, centromere function	transposase	eukaryotes	(SMIT and RIGGS 1996)
Pdc2/Rag3	transcription factor, regulation of pyruvate utilization	transposase	yeast	(HOHMANN 1993)
Jerky	neuron mRNA-binding protein, mutated in epilepsy syndromes	transposase	mammals	(TOTH *et al.* 1995)
Jerky-like	unknown	transposase	mammals	(ZENG *et al.* 1997)
KIAA0461	unknown	transposase	mammals	(SMIT 1999)
Tigger-derived	unknown	transposase	mammals	(ROBERTSON 2002)
Metnase	DNA integration, radiation resistance, DNA break repair	transposase	human	(LEE *et al.* 2005)

Table 1.2 continued

Gene	Protein functions	TE gene	Organisms	References
P-derived	DNA binding protein	transposase	*Drosophila*	(PINSKER *et al.* 2001; REISS *et al.* 2005)
Phsa/Pgga	unknown	transposase	mammals, birds	(HAMMER *et al.* 2005)
Pgbd1-4	unknown	transposase	human	(SARKAR *et al.* 2003)
Pgbd5	unknown	transposase	bony vertebrates	(SARKAR *et al.* 2003)
HarbI1	unknown	transposase	bony vertebrates	(KAPITONOV and JURKA 2004)
Fhy3, Far1	transcription factors, sensitivity to far-red light	transposase	*Arabidopsis*	(HUDSON *et al.* 2003)
Mustang	unknown	transposase	flowering plants	(COWAN *et al.* 2005)

Impact of transposable elements on the evolution of host gene regulation

The ability of TEs to alter the regulation and expression patterns of host genes is well-documented (BI *et al.* 1997; DUNN *et al.* 2003; JORDAN *et al.* 2003; LANDRY *et al.* 2002; MEDSTRAND *et al.* 2001; VAN DE LAGEMAAT *et al.* 2003) and has been discussed in a number of reviews (BROSIUS 1999; HAMDI *et al.* 2000; KIDWELL and LISCH 1997; KIDWELL and LISCH 2001; TOMILIN 1999). TE sequences are saturated by transcription factor binding sites and serve as transcriptional promoters, enhancers or silencers for nearby genes (BANVILLE and BOIE 1989; BRITTEN 1997; BROSIUS 1999; FRIESEN *et al.* 1986; HEWITT *et al.* 1995; KAZAKOV and TOMILIN 1996; YANG *et al.* 1998) (Table 1.3). Retroelements can also act as alternative promoters (MEDSTRAND *et al.* 2001), bidirectional promoters (DOMANSKY *et al.* 2000) or may compete with gene promoters for the binding of transcription factors (CONTE *et al.* 2002). Many intriguing cases where TE-derived promoters contribute to tissue-specific gene expression were shown (VAN DE LAGEMAAT *et al.* 2003). For instance, many LTR promoters and enhancers are active primarily in the placenta (BANVILLE and BOIE 1989; BI *et al.* 1997; BIECHE *et al.* 2003;

CHANG-YEH *et al.* 1991; LANDRY *et al.* 2002; MEDSTRAND *et al.* 2001; SCHULTE *et al.* 1996). In addition to 3' UTR AU-rich elements that regulate mRNA stability (BRITTEN 2006), retroposons such as Alu can provide polyadenylation signals to host genes. Some Alu insertions contain transcriptional regulatory sequences, such as a retinoic-acid-response element (VANSANT and REYNOLDS 1995). Both LTR and other TEs contain potential hormone-responsive sites (BABICH *et al.* 1999; RAMAKRISHNAN and ROBINS 1997) which have been implicated in hormone-dependent regulation of several human genes (NORRIS *et al.* 1995; VANSANT and REYNOLDS 1995). Furthermore, TEs can control genes epigenetically when inserted within or very close to the genes (LIPPMAN *et al.* 2004) by two mechanisms: directly by inducing the methylation (and therefore silencing) of neighboring DNA and indirectly by disrupting the normal epigenetic state of a nearby gene.

Table 1.3: Vertebrate regulatory elements generated by TEs [information from (BROSIUS 1999; MEDSTRAND *et al.* 2005)]

TE	Gene	Organism	Function	References
Alu	θ1 globin	Higher primates	CCAAT box of promoter	(KIM *et al.* 1989)
Alu	Myeloperoxidase	Human	Composite SP1-thyroid hormone-retinoic acid response element	(PIEDRAFITA *et al.* 1996)
Alu	BRCA-1 gene, ERF-3	Human	Estrogen-dependent transcriptional enhancers	(NORRIS *et al.* 1995)
Alu	Parathyroid hormone gene	Human	Negative calcium response element	(MCHAFFIE and RALSTON 1995)
Alu	Haptoglobin related gene	Human	Transcriptional enhancer	(OLIVIERO and MONACI 1988)
Alu	Adenosine deaminase	Human	Transcriptional enhancer	(ARONOW *et al.* 1992)

Table 1.3 continued

TE	Gene	Organism	Function	References
Alu	Mitochondrial hinge protein	Human	Transcriptional enhancer	(LIU and BRADNER 1993)
Alu	SV40 origin	Human	Transcriptional enhancer	(SAEGUSA *et al.* 1993)
Alu	CD8α	Human	Transcriptional enhancer	(HAMBOR *et al.* 1993)
Alu	Keratin 18 (human)	Mouse (transgenic)	Transcriptional insulation; Alus provide retinoic acid receptor binding sites	(NEZNANOV and OSHIMA 1993; THOREY *et al.* 1993; VANSANT and REYNOLDS 1995)
Alu	α3 nicotinic receptor subunit	Human	Transcriptional modulation	(FORNASARI *et al.* 1997)
Alu	ε-globin	Human	Transcriptional modulation	(WU *et al.* 1990)
Alu	c-myc	Human	Transcriptional modulation	(TOMILIN *et al.* 1990)
Alu	Potentially many genes	Human	Transcriptional modulation via binding of YY1 protein	(HUMPHREY *et al.* 1996)
Alu	Poly(ADP-ribosyl) transferase (ADPRT) gene	Human	Transcription regulation	(OEI *et al.* 1997; SCHWEIGER *et al.* 1995)
Alu	$Fc\varepsilon RI_{\gamma}$	Human	Transcriptional regulation (positive and negative)	(BRINI *et al.* 1993)
Alu	Proliferating cell nuclear antigen (PCNA)	Human	Transcriptional silencer	(SELL *et al.* 1992)
Alu	Wilms tumour suppressor gene (WT1)	Human	Transcriptional silencer	(HEWITT *et al.* 1995)
Alu	interferon-γ	Human	Transcription factor binding site	(ACKERMAN *et al.* 2002)
Alu	PAX6	Human	Transcription factor binding site	(ZHOU *et al.* 2002)
B1	Immunoglobulin κ light chain	Mouse	Negative regulation of transcription	(SAKSELA and BALTIMORE 1993)

Table 1.3 continued

TE	Gene	Organism	Function	References
B2	MOK-2 zinc-finger protein	Mouse	Exerts a negative cis-acting effect on *MOK-2* promoter activity	(ARRANZ *et al.* 1994)
B2	Fourth component of complement (C4) in H-2^k haplotype	Mouse	Reduces expression rate to 1/10 of non-H-2^k mice	(ZHENG *et al.* 1992)
B2	MHC class I genes	Mouse	Polyadenylation signal	(KRESS *et al.* 1984)
B2	$B2^+$ $mRNA_x$	Mouse	Polyadenylation signal	(RYSKOV *et al.* 1984)
B2	Glutathione S-transferase	Mouse	Polyadenylation signal	(ROTHKOPF *et al.* 1986)
B2	Muscle γ-phosphorylase kinase	Mouse	Polyadenylation signal	(MAICHELE *et al.* 1993)
CR1	Lysozyme	Chicken	Transcriptional silencer	(BANIAHMAD *et al.* 1987)
HERV-E	Salivary amylase gene	Human	Promoter	(EMI *et al.* 1988; SAMUELSON *et al.* 1990; TING *et al.* 1992)
L1	Thymidylate synthase	Mouse	Polyadenylation signal	(HARENDZA and JOHNSON 1990)
L1	Proteasome activator PA28β (PMSE2b)	Mouse	Promoter	(ZAISS and KLOETZEL 1999)
L1	Apolipoprotein (a)	Human	Transcriptional enhancer	(YANG *et al.* 1998)
L1	Insulin I gene	Rat	Transcriptional silencer	(LAIMINS *et al.* 1986)
L2	Annexin VI, interleukin-4, protein kinase C-β	Human	T-cell specific silencer	(DONNELLY *et al.* 1999)
LTR	Leptin	Human	Placental enhancer	(BI *et al.* 1997)
LTR	cDNA 7, $cDNA_\gamma$	Human	Polyadenylation signal	(PAULSON *et al.* 1987)
LTR	PLT	Human	Polyadenylation signal	(GOODCHILD *et al.* 1992)
LTR	cH-6	Human	Polyadenylation signal	(MAGER 1989)

Table 1.3 continued

TE	Gene	Organism	Function	References
LTR	cH-7	Human	Polyadenylation signal	(MAGER 1989)
LTR	cPB-3	Human	Polyadenylation signal	(MAGER 1989)
LTR	carbonic anhydrase *(CA1)*	Human	Erythroid-specific promoter	(VAN DE LAGEMAAT *et al.* 2003)
LTR	endothelin-B receptor	Human	Trophoblast-specific promoter	(LANDRY *et al.* 2003; MEDSTRAND *et al.* 2001)
LTR	Mid1	Human	Tissue-specific promoter (fetal kidney and placenta)	(LANDRY *et al.* 2002)
LTR	pleiotrophin	Human	Tissue-specific promoter (placenta)	(SCHULTE *et al.* 1996)
LTR	insulin-like peptide INSL4	Human	Tissue-specific promoter (placenta)	(BIECHE *et al.* 2003)
LTR	Oncomodulin	Rat	Promoter	(BANVILLE and BOIE 1989)
LTR	MIPP	Mouse	Promoter	(CHANG-YEH *et al.* 1991)
LTR	AF-3	Human	Promoter	(FEUCHTER *et al.* 1992)
LTR	AF-4 (CDC4L homology)	Human	Promoter	(FEUCHTER *et al.* 1992)
LTR	PLA2L (phospholipase A2 homology)	Human	Promoter	(FEUCHTER-MURTHY *et al.* 1993)
LTR	Calibindin D28K	Human	Promoter	(LIU and ABRAHAM 1991)
LTR	ZNF80 zinc finger gene	Human	Promoter	(DI CRISTOFANO *et al.* 1995)
LTR	Aromatase	Chicken	Promoter	(MATSUMINE *et al.* 1991)
LTR	ß1,3-galactosyltransferase 5 (ß3Gal-T5)	Human	Promoter	(DUNN *et al.* 2003)

Table 1.3 continued

TE	Gene	Organism	Function	References
LTR	Sex-limited protein (slp)	Mouse	Promoter	(RAMAKRISHNAN and ROBINS 1997; ROBINS and SAMUELSON 1992; STAVENHAGEN and ROBINS 1988)
LTR	apolipoprotein C-I	Human	Promoter	(LANDRY *et al.* 2003; MEDSTRAND *et al.* 2001)
LTR	alcohol dehydrogenase 1C	Human	cis-acting element	(CHEN *et al.* 2002)
LTR	BAAT gene	Human	Promoter	(VAN DE LAGEMAAT *et al.* 2003)
LTR	aromatase CYP19	Human	Promoter	(VAN DE LAGEMAAT *et al.* 2003)
LTR	locus control region in the human ß-globin gene cluster	Human	Promoter	(PLANT *et al.* 2001)
LTR-IS	A1	Mouse	Polyadenylation signal	(BAUMRUKER *et al.* 1988)
LTR-IS	A3	Mouse	Polyadenylation signal	(BAUMRUKER *et al.* 1988)
MIR	β-tubulin	Human	Polyadenylation signal	(MURNANE and MORALES 1995)
MIR	Follitropin receptor	Sheep	Polyadenylation signal	(MURNANE and MORALES 1995)
MIR	Clone c-zrog02	Human	Polyadenylation signal	(MURNANE and MORALES 1995)
MIR	Clone NIB1273	Human	Polyadenylation signal	(MURNANE and MORALES 1995)

Two recent studies have estimated the impact of TEs on host gene regulation on a genome-wide scale. 25% of all known human genes (JORDAN *et al.* 2003; VAN DE

LAGEMAAT *et al.* 2003) and mouse genes (VAN DE LAGEMAAT *et al.* 2003) contain TEs within their UTR and/or promoter regions. Approximately 8% of all proximal promoter regions and 2.5% of known transcription factor binding sites of human genes were located within a TE (JORDAN *et al.* 2003). Besides these cis-regulating effects, more than 50% of scaffold/matrix attachment regions (S/MARs) are made up of TEs (JORDAN *et al.* 2003). These results suggested that TEs may have a substantial impact on the evolution of human gene regulation, as was first hypothesized by Britten and Davidson (BRITTEN and DAVIDSON 1969), before the connection between TEs and repetitive DNA was established.

Due to the variety of TE functions related to gene regulation, there is a large possibility that more are still uncovered. As recently emphasized, the majority of human transcripts do not encode proteins (CLAVERIE 2005). These non-coding RNA (ncRNA) can function directly as structural, catalytic or regulatory RNAs (MATTICK and MAKUNIN 2006). Several different systematic investigations have recently identified an unexpectedly large number of ncRNA genes which can potentially be the large source of novel gene regulators (HUTTENHOFER and VOGEL 2006; MARKER *et al.* 2002; STORZ 2002; WASHIETL *et al.* 2005a; WASHIETL *et al.* 2005b; WASSARMAN *et al.* 2001). The availability of these new data opens the opportunity for exploring the relationship between non-coding regulatory genes and TEs.

Introduction to microRNAs (miRNAs) and small interfering RNAs (siRNAs)

Currently, considerable evidence indicates that small noncoding RNAs can play a major role in regulating gene expression in eukaryotes (CULLEN 2002; HUTVAGNER and ZAMORE 2002b). Of particular interest are a class of ~22-nt RNAs that can be divided

into small interfering RNAs (siRNAs) and microRNAs (miRNAs) (AMBROS *et al.* 2003). siRNAs are derived from long, double-stranded RNAs which are processed into shorter duplexes by Dicer ribonuclease (BERNSTEIN *et al.* 2001; KNIGHT and BASS 2001), and then one strand of the duplex is incorporated into RNA induced silencing complex (RISC) (AMBROS *et al.* 2003; MARTINEZ *et al.* 2002; SCHWARZ *et al.* 2002). The siRNA component guides RISC to mRNA molecules containing a homologous antisense sequence, resulting in degradation of that mRNA (MARTINEZ *et al.* 2002; SCHWARZ *et al.* 2002). This process is termed RNA interference (RNAi) (FIRE *et al.* 1998) which is thought to have originally evolved to silence viruses and TEs (BUCHON and VAURY 2006).

Similar to siRNAs, miRNAs are ~22nt single-stranded noncoding RNAs that regulate the expression of complementary mRNAs (BARTEL 2004). In contrast to siRNAs, miRNAs are derived by processing of a ~70-nt RNA stem–loop (hairpin) structure termed a pre-miRNAs (AMBROS *et al.* 2003; LAGOS-QUINTANA *et al.* 2001; LAU *et al.* 2001; LEE and AMBROS 2001). In animals, pre-miRNAs are transcribed as longer primary transcripts (pri-miRNAs) that are processed by Drosha in the nucleus into compact, folded structures (pre-miRNAs), then exported to the cytoplasm, where they are cleaved by Dicer to yield mature miRNAs (LEE *et al.* 2002) and are incorporated into a ribonucleoprotein complex (MOURELATOS *et al.* 2002).

As regulators of gene expression, miRNAs can work by basically two modes (DOENCH *et al.* 2003; HUTVAGNER and ZAMORE 2002a; RHOADES *et al.* 2002; TANG *et al.* 2003; ZENG *et al.* 2003). miRNAs can act through translational repression of their target mRNAs and may also cause mRNA degradation of their target genes via an RNAi-

like mechanism (HUTVAGNER and ZAMORE 2002a; LLAVE *et al.* 2002; YEKTA *et al.* 2004; ZENG *et al.* 2003). Recently, anti-correlated expression patterns between miRNA sequences and their target mRNAs have provided evidence in favor of the mRNA degradation model (HUANG *et al.* 2006). Hundreds of miRNA genes have been found in animals, and the majority of these are phylogenetically conserved (AMBROS 2004). Their importance is supported by the many biological processes in which they are participated, including developmental timing, cell proliferation, apoptosis, metabolism, cell differentiation, and morphogenesis (ALVAREZ-GARCIA and MISKA 2005; AMBROS 2004).

One previously recognized distinction between miRNAs and siRNAs is that miRNAs are usually found in introns and intergenic regions (BARTEL 2004), while siRNAs originate from within genes and TEs (MATZKE *et al.* 2000; SLOTKIN *et al.* 2005; VASTENHOUW and PLASTERK 2004). Interestingly, the relationship to TEs has been pointed out as a discrepancy between miRNAs and siRNAs, which are closely related in terms of structure, function, and biogenesis. A major goal of my dissertation research was to address whether or not the TE-based distinction between siRNAs and miRNAs was justified and valid. miRNAs are estimated to comprise 1%–5% of animal genes (BARTEL 2004; BENTWICH *et al.* 2005; BEREZIKOV *et al.* 2005), making them one of the most abundant classes of regulators. The growing number of annotated miRNA genes, enhanced by advanced experimental techniques and new detection programs, has provided an opportunity to explore the relationship between TEs and evolution of these new regulatory elements as well as the possibility of an evolutionary connection between siRNAs and miRNAs through TEs.

Analysis of origin and evolution of gene sequences derived from TEs

This dissertation focuses exclusively on eukaryotic TEs, with an emphasis on human TEs that have contributed to the evolution of protein coding sequences (CHAPTER 2 and 3) and non-coding regulatory RNAs (CHAPTER 4, 5 and 6).

CHAPTER 2 presents a detailed analysis of exonization events of LTR elements in the human genome. The distribution patterns of LTR retrotransposon sequences in human exons were determined. Using new experimental data, the evolutionary history of the exonization process of an alternatively spliced exon of *IL22RA2* was reconstructed.

CHAPTER 3 illustrates the ability of different classes of sequence similarity search methods to detect TE-derived sequences in experimentally characterized proteins. The ascertainment biases related to these search methods were evaluated. The probabilistic analysis of TE-derived exon sequences was applied to determine the potential of TEs, particularly non-coding TEs, to contribute protein coding sequences to human genome.

CHAPTER 4 determines the extent of TE contributions to human miRNA genes along with the evolutionary dynamics of TE-derived human miRNAs. The potential regulatory and functional significance of TE-derived miRNAs was explored by combining information on miRNA target site prediction, expression data for miRNA-mRNApairs, and gene functional annotations. An *ab initio* prediction algorithm I developed was used to discover putative cases of novel TE-derived miRNA genes.

CHAPTER 5 demonstrates the investigation of a recently discovered family of human miRNA genes, hsa-mir-548, which was found in this study to be derived from

Made1 TEs. The analysis of hsa-mir-548 target genes in terms of gene expression and functional affinities indicates a potential role for this miRNA family in cancer.

CHAPTER 6 proposes a specific model whereby miRNAs encoded from short non-autonomous DNA-type TEs, known as MITEs, evolved from corresponding ancestral autonomous elements that originally encoded siRNAs. A computational analysis of genome sequences, annotation and expression data from the plants *Arabidopsis thaliana* and *Oryza sativa* (rice) was performed to predict the dual coding siRNA-miRNA TEs, which represent evolutionary intermediates in the transition from siRNAs to miRNAs.

CHAPTER 2

EXONIZATION OF THE LTR TRANSPOSABLE ELEMENTS IN HUMAN GENOME

ABSTRACT

Background

Retrotransposons have been shown to contribute to evolution of both structure and regulation of protein coding genes. It has been postulated that the primary mechanism by which retrotransposons contribute to structural gene evolution is through insertion into an intron or a gene flanking region, and subsequent incorporation into an exon.

Results

We found that Long Terminal Repeat (LTR) retrotransposons are associated with 1,057 human genes (5.8%). In 256 cases LTR retrotransposons were observed in protein-coding regions, while 50 distinct protein coding exons in 45 genes were comprised exclusively of LTR RetroTransposon Sequence (LRTS). We go on to reconstruct the evolutionary history of an alternatively spliced exon of the Interleukin 22 receptor, alpha 2 gene (*IL22RA2*) derived from a sequence of retrotransposon of the Mammalian apparent LTR retrotransposons (MaLR) family. Sequencing and analysis of the homologous regions of genomes of several primates indicate that the LTR retrotransposon was inserted into the *IL22RA2* gene at least prior to the divergence of Apes and Old World monkeys from a common ancestor (~ 25 MYA). We hypothesize

that the recruitment of the part of LTR as a novel exon in great ape species occurred prior to the divergence of orangutans and humans from a common ancestor (~ 14 MYA) as a result of a single mutation in the proto-splice site.

Conclusions

Our analysis of LRTS exonization events has shown that the patterns of LRTS distribution in human exons support the hypothesis that LRTS played a significant role in human gene evolution by providing cis-regulatory sequences; direct incorporation of LTR sequences into protein coding regions was observed less frequently. Combination of computational and experimental approaches used for tracing the history of the LTR exonization process of *IL22RA2* gene presents a promising strategy that could facilitate further studies of transposon initiated gene evolution.

INTRODUCTION

Retrotransposon sequences comprise more than 40 % of the human genome (JASINSKA and KRZYZOSIAK 2004; LANDER *et al.* 2001). Once dismissed as "junk DNA" of little or no adaptive significance (DOOLITTLE and SAPIENZA 1980; OHNO 1972), retrotransposons and other classes of transposable elements (TEs) are now generally considered as significant contributors to gene and genome evolution (BROSIUS 1999; BROSIUS and GOULD 1992; KAZAZIAN 2004; KIDWELL and LISCH 2001; MAKALOWSKI 2003). Of particular interest has been the ability of TEs to contribute to exon evolution by "exonization", *i.e.*, an insertion of a TE into an intron and subsequent recruitment of this sequence or its part into a new protein-coding exon (NEKRUTENKO and LI 2001). For example, it has been estimated that 5% of all alternatively spliced human exons had been

derived from the exonization of Alu elements (DAGAN *et al.* 2004; MAKALOWSKI *et al.* 1994; SOREK *et al.* 2002).

LTR transposable elements comprise nearly one-tenth of the human genome and have been implicated in the cis-regulatory evolution of a number of human genes (BI *et al.* 1997; BROSIUS 1999; BROSIUS and GOULD 1992; DUNN *et al.* 2003; LANDRY *et al.* 2002; MEDSTRAND *et al.* 2001; VAN DE LAGEMAAT *et al.* 2003). The structure of a complete LTR retrotransposon (autonomous mobile element) comprises two copies of long terminal directed repeats (LTRs) flanking an internal region containing *gag* and *pol* genes, which encode a protease, reverse transcriptase, RNase H and integrase. These protein products are necessary for the formation of virus-like particles (VLPs) wherein replication of the element takes place. Some elements evolved from retroviruses have additional open reading frames (ORFs), e.g. *env* gene (LANDER *et al.* 2001; SEMIN and IL'IN IU 2005). Flanking LTRs contain all the necessary transcriptional regulatory elements.

Although global database screens have been conducted to examine the contribution of TEs to human protein-coding regions (BRITTEN 2006; LORENC and MAKALOWSKI 2003; NEKRUTENKO and LI 2001), none have concentrated specifically on the prevalence of the LRTS-derived protein-coding exons of human genes. Here we report the results of computational analysis of the LRTS exonization in human genome. Also we describe the plausible scenario of the exonization process of an alternatively spliced exon of the alpha 2 gene of the Interleukin 22 receptor (*IL22RA2*) supported by new experimental data.

RESULTS AND DISCUSSION

Updated list of LRTS-associated genes

To identify incidences of LRTS exonization, the annotation of human exons given in the UCSC genome browser was compared with the annotation of transposable elements available in the same source. We detected LRTS associations in 1,057 out of 18,241 genes (5.8 %). These associations include 1,249 distinct exons participating in 1,287 transcripts (note that a particular exon is counted once though it may participate in several alternative transcripts). It was reported earlier (NEKRUTENKO and LI 2001) that 130 out of 13,799 human genes (0.9 %) were found to contain LRTS in protein coding regions. In comparison, in our data set (18,241 genes/23,821 transcripts) we observed LRTS associations with protein-coding exons in 256 genes (1.4 %). Current LRTS search done at the DNA instead of mRNA level helped detect several short LRTS-exon overlaps that could be missed at mRNA level. Interestingly, only 53 of the previously reported 130 cases were found in current analysis using the updated RefSeq gene data. Many previously identified cases (61 cases) did not show up in our data set as the earlier sequences were removed, suppressed, or replaced. Several cases appear to be possible false positives. In one case, LRTS was detected in UTR instead of in CDS. No LRTS was detected in other two cases when the RepeatMasker program was run separately on each mRNA sequence using its specific G+C content, which gives a slightly more accurate result, as opposed to input of multiple sequences with averaged G+C content used in the program (http://biowulf.nih.gov/apps/repeatmasker/repeatmasker_help.txt).

Distribution of LRTS in human exons

We found that human gene exons (either protein-coding or non-coding) overlap with LTR flanks of LTR elements more frequently (1,074 cases) than with internal sequences (242 cases; note that exons overlaped with both regions were counted twice). This observation could be related to the fact that most (85%) of the LTR retroposon-derived sequences in human genome consist only of a solo LTR, with the internal sequence lost due to homologous recombination between the flanking LTRs (LANDER *et al.* 2001). Upon checking by BLASTX of 242 exons overlapping with the internal sequences, 61 exons were found to contain a section or even a whole viral gene (*i.e. gag*, *pol*, and *env*). However, only 20 of these 61 exons were protein-coding exons. Moreover, only in 10 cases was the reading frame of a human gene the same as the one of the viral gene. Seven out of these ten cases were observed in hypothetical genes. The remaining three cases represented a gene for endogenous retroviral protein, syncytin (*ERVWE1*), a gene for Krueppel-related zinc finger protein (*H-plk*) and a placenta-specific gene (*PLAC4*) which protein products contain the envelope, envelope and gag viral protein domain, respectively. All three genes are preferentially expressed in the placenta (BLOND *et al.* 1999; KATO *et al.* 1990; KIDO *et al.* 1993). This observation indicates that the invasion of the Human Endogenous Retrovirus (HERV) may contribute to molecular mechanisms involved in human reproduction (MUIR *et al.* 2004).

The majority of exons overlapping with LRTS (1,123 of 1,249) contain sequences homologous to only one LRTS. Exons overlapping with more than one LRTS were observed as well (Table 2.1). Overall, we have found 1,395 associations (overlaps) between an LRTS and an exonic sequence. These 1,395 observations were classified

further according to the extent of LRTS overlap with an exon (Table 2.2), type of exon (Table 2.3), and LRTS class/family (Table 2.4). The majority of LRTS associations with genes (586/1395 or 42 %) constitute an apparent extension of original exon due to activation of alternative splice site located inside LRTS. On the other hand, in 22.9% (319/1395) of these associations LRTS was recruited as an entirely novel exon (Table 2.2).

Regarding the distribution of LRTS within a complete gene structure (5'UTR, first CDS exon, internal protein coding exons, last CDS exon, 3'UTR), the LRTS fragments were found in untranslated regions (UTRs), mainly in 3'UTRs, much more frequently than in protein-coding (CDS) regions. This observation is consistent with the previous study (JORDAN *et al.* 2003) and indicates the putative role of LRTS in resident gene regulation by providing sequence material for emerging regulatory sequences (BROSIUS 1999; MEDSTRAND *et al.* 2001). In comparison, insertion of LRTS in a protein coding region may interfere with gene function, and in many cases such a modification is likely to be eliminated by negative selection. Note that an LRTS was found more frequently in the last CDS exon, especially in the exon untranslated region, and less frequently in internal coding exons (Table 2.3).

Table 2.1: The distribution of the number of LTR elements (either partial or full elements) containing in an exon

number of LTR elements overlaps with an exon	number of exons	number of associations
1	1123	1123
2	108	216
3	16	48
4	2	8
Total	1249	1395

Table 2.2: The distribution of the extent of overlap between an exon and an LTR element

extent of LRTS overlap	number of associations
An LRTS completely covers an exon	319
An LRTS partially overlaps (5' or 3' boundaries) with an exon	586
An LRTS is situated within an exon	490
Total	1395

Table 2.3: The distribution of type of exons containing LRTS

exon type	number of associations
5'UTR exon	245
3'UTR exon	196
First CDS exon	127 (41 in the 5'UTR, 8 in the CDS region and 78 span both regions)
Last CDS exon	571 (484 in the 3'UTR, 16 in the CDS region and 71 span both regions)
Single protein coding exon	152 (17 in the 5'UTR, 97 in the 3'UTR, 11 in the CDS region and 27 span both UTR and CDS regions)
Internal protein coding exon	72
More than one type of exon (for a particular exonic sequence)	32
Total	1395

Table 2.4: The distribution of class/family of LRTS containing in an exon

LRTS class/ family	number of associations
ERV1	513
ERVK	58
ERVL	249
MaLR	575
Total	1395

LRTS-derived protein coding exons

We have found 50 protein coding exons completely derived from LRTS (41 internal, 2 initial, 4 terminal coding exons and 3 single coding exons (Table A.1). Most of LRTS-derived exons (36/50) were comprised exclusively of LTR flanking regions. Eleven exons were derived from LTR element internal sequences and 3 exons contained both types of regions. Of the 50 exons, 38 were components of well characterized protein coding genes (*i.e.*, genes with the corresponding mRNAs available in GenBank and with encoded proteins listed in SWISS-PROT, TrEMBL, and TrEMBL-NEW).

The low frequency of protein coding exons fully derived from LRTS indicates that the chance of a successful recruitment of a whole coding exon from the LTR transposable element is rather small. The exonization of originally intronic LRTS requires the presence of a pair of potential splice sites, enclosing a sequence with no stop codon in the appropriate reading frame. Also, the amino acids contributed by a mobile element should not disrupt the structure of a protein encoded by the original gene, particularly, the addition of a new exon should not change the coding frame for the remaining part of a gene.

Interestingly, most of the protein coding exons derived entirely from the LTR flanking regions originated from the MaLR family (24 out of 36). This could be explained by several factors. First of all, MaLR elements make up about 50% of the LTR retroelements in the human genome (LANDER *et al.* 2001), and this high frequency alone may relate to their over-representation in protein coding exons. MaLRs are also relatively ancient elements, which have probably been exposed to more opportunities for exonizations over time. Note that the age factor has been implicated for proliferation of

Alu-derived exons as well (SOREK *et al.* 2002). Finally, it is a formal possibility that nucleotide sequences of the MaLR family are better amenable for derivation of protein coding exons. The internal sequence of MaLR is rarely found retained in the human genomic sequence (SMIT 1993). Particularly, among exons derived from the internal parts of LRTS only one was from the MaLR family.

Contribution of LRTS to gene transcripts

We further analyzed the abundance of LRTS-derived exons in gene transcripts. Most of the 275 genes containing at least one exon completely derived from LRTS (201 out of 275) are single transcript genes while the remaining 74 generate more than one transcript per gene. Note that about 60% (121/201) of single transcript genes encode zinc finger proteins (25%) or hypothetical proteins (35%). Apparently for the single transcript gene the LRTS insertion either has not disrupted the host gene function or possibly provided some beneficial modulation of the initial function and thus has been tolerated by natural selection.

In 55 out of 74 genes (74.3%) with multiple transcripts, LRTS-derived exons were present in some transcript variants, but not in all of them. This observation corresponds to the scenario whereby recruiting of LRTS into alternatively spliced exon allows the main transcript to maintain the function while the LRTS-associated exons are "examined" by natural selection, which may lead to emergence of transcripts with new functions.

We also found that most of the LRTS-derived protein coding exons (48/50) were either alternatively spliced ones or the components of single transcript genes. In contrast, most of LRTS derived constitutive exons (those that are present in all alternative

transcripts) are found in 5' UTR sequences. This observation indicates that novel cis-regulatory sequences supplied by LTR elements to human genes are more likely to be fixed in evolution than sequences supplying protein coding domain which are used as alternative ways to create protein variability.

Reconstruction of evolution of *IL22RA2* gene (transcript variant 1)

The *IL22RA2* gene has an internal protein coding exon derived from an LTR flanking sequence. This gene encodes the only soluble receptor (WEISS *et al.* 2004) in the class II cytokine receptor family (CRF2). IL22RA2 protein specifically binds to interleukin 22 (IL22) and by preventing the interaction of IL22 with its cell surface receptor, neutralizes IL22 activity (DUMOUTIER *et al.* 2001; KOTENKO *et al.* 2001; XU *et al.* 2001). Three alternatively spliced transcripts of the *IL22RA2* human gene encoding three protein variants (263, 231 and 130 amino acids in length) have been described earlier (KOTENKO *et al.* 2001). The longest transcript (variant 1) is generated (Figure 2.1) by addition of the 96 nt exon (exon 3/4) to splice variant 2 between exon 3 and exon 4 (DUMOUTIER *et al.* 2001; GRUENBERG *et al.* 2001; KOTENKO *et al.* 2001).

In the current study, we provide experimental data and computational analysis that show evolutionary evidence of exonization of LRTS invaded the human *IL22RA2* gene. The exon 3/4 of the *IL22RA2* gene (transcript variant 1) is situated within the LTR sequence of MSTB2 subfamily of MaLR family (found in the same orientation as the coding sequence (Figure 2.1)). The sequence alignment of the particular LTR and the MSTB2 LTR consensus sequence shows 82.8 % identity (for ungapped part of the 431 nt long alignment). The exon 3/4 contributes 32 amino acids to the IL22RA2 protein product without changing reading frame for the rest of the protein. A homologous exon

was not found either in the mouse or in the rat orthologous gene. Weiss et al. 2004 (WEISS *et al.* 2004) also indicated that a counterpart of this exon was absent in mouse and rat. The functionality of the LTR exonization is corroborated by the existence of the mRNA sequences containing the exon 3/4 [RefSeq: NM_052962, GenBank: AY040567, AY358737, EMBL: AJ313162]. The data available at the UCSC genome browser show that the MSTB2 derived sequence is conserved in chimpanzee and rhesus monkey while is absent in other vertebrates. To extract the sequences homologous to the exon 3/4 in seven primates: human, chimpanzee, bonobo, gorilla, orangutan, crab-eating macaque and rhesus monkey, we have performed the PCRs with human DNA derived primers (see methods), which generated well interpretable PCR products for all species (Figure 2.2). We used newly determined PCR product sequences as well as publicly available genomic sequences of human, chimpanzee and rhesus monkey to construct the multiple sequence alignment. We observed that the splice sites flanking the target exon in all species but the crab-eating macaque and the rhesus monkey followed the GT/AG rule. In the other two species, we observed AT instead of GT at the donor site (Figure 2.3). Therefore, emergence of this exon was likely to occur in ape lineage earlier than the divergence of orangutans and humans (Figure 2.4). This event was mediated by the single transition from A to G yielding canonical donor splice site consensus. Note that AT (or GT in other cases) is positioned in the predicted LTR polyadenylation site A**AT**AAA (Figure 2.3). Contrary to the acceptor site, the strength of the donor site depends on the presence of just a few specific nucleotides around GT consensus. Therefore, a single mutation might create a functional donor splice site. The canonical dinucleotide (AG) of the acceptor site appeared in all primates we have studied. However, this dinucleotide is different from

dinucleotide (GC) situated in the same position in MSTB2 consensus sequence (Figure 2.3). One possibility is that the mutation of GC to AG could happen earlier in the primate lineage. However, the sequence logo generated from the multiple sequence alignment of the 880 MSTB2 sequences existing in the human genome shows low degree of conservation in the vicinity of acceptor site. Therefore, the dinucleotide predecessor of AG should not necessarily be the consensus GC dinucleotide.

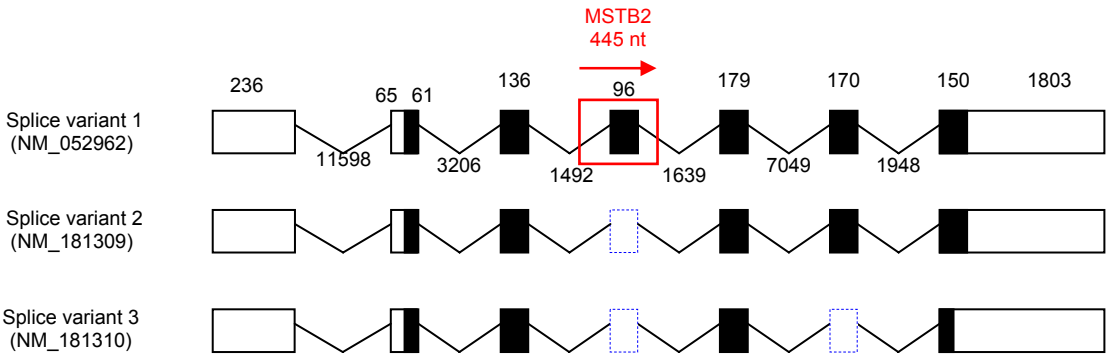

Figure 2.1: Exon-intron organization of human *IL22RA2* gene. Exon and intron sequences are represented by boxes and angular lines, respectively, with lengths indicated in base pairs. Coding and untranslated regions are represented by filled boxes and open boxes, respectively while the blue dashed boxes demonstrate the absence of the exon sequences on mRNA level. The region showing homology with MSTB2 is labeled in red border. A horizontal arrow indicates the LTR orientation.

Several coincidences must have been involved in creation of the exon 3/4. The viable structural elements of the splice sites (GT/AG) were created by mutations. With the upstream intron in phase 2, the exon 3/4 emerged in the frame which had no stop codons inside, while the other two possible phases of intron would cause premature termination of translation. The new exon 3/4 (with length divisible by three) did not disrupt the global reading frame and therefore did not change the downstream amino acid sequence known to be important for ligand binding (GRUENBERG *et al.* 2001). Our

findings show that the exon 3/4 of *IL22RA2* might be active and be expressed in the Great Apes, while we have not confirmed its expression in the Old World monkeys. This observation indicates that the exon 3/4 is likely to possess functional properties and it is an alternatively spliced exon. We have evaluated the possibility that the exon 3/4 is the subject for positive selection by the standard test based on non-synonymous K_a to synonymous K_s divergence rates ratio. There are three nonsynonymous substitutions between human and orangutan homologous exonic sequences, while there are no synonymous substitutions. The use of the Laplace pseudocounts produces $(K_a+1)/(K_s+1) > 1$, which indicates possible positive selection.

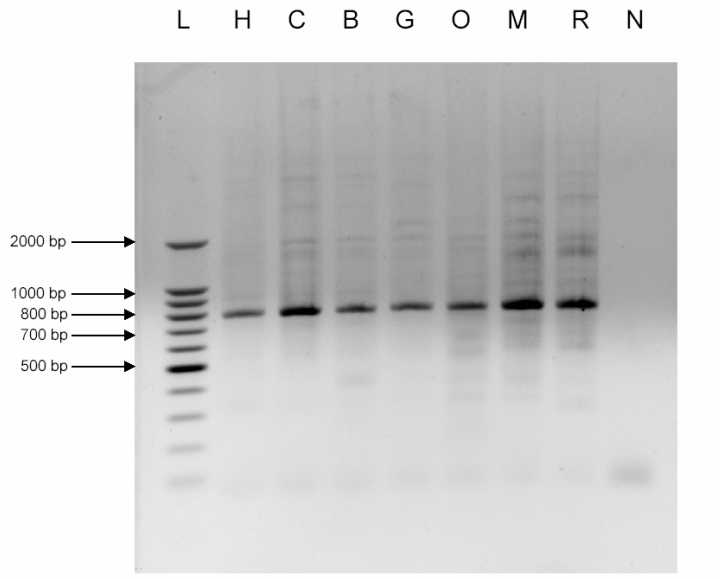

Figure 2.2: PCR-sequencing. Agarose gel electrophoresis of *IL22RA2* homologous regions carrying LTR, MSTB2, from seven primates amplified by PCR. L, ladder; H, human; C, chimpanzee; B, bonobo; G, gorilla; O, orangutan; M, crab eating macaque; R, rhesus monkey; N, nontemplate control.

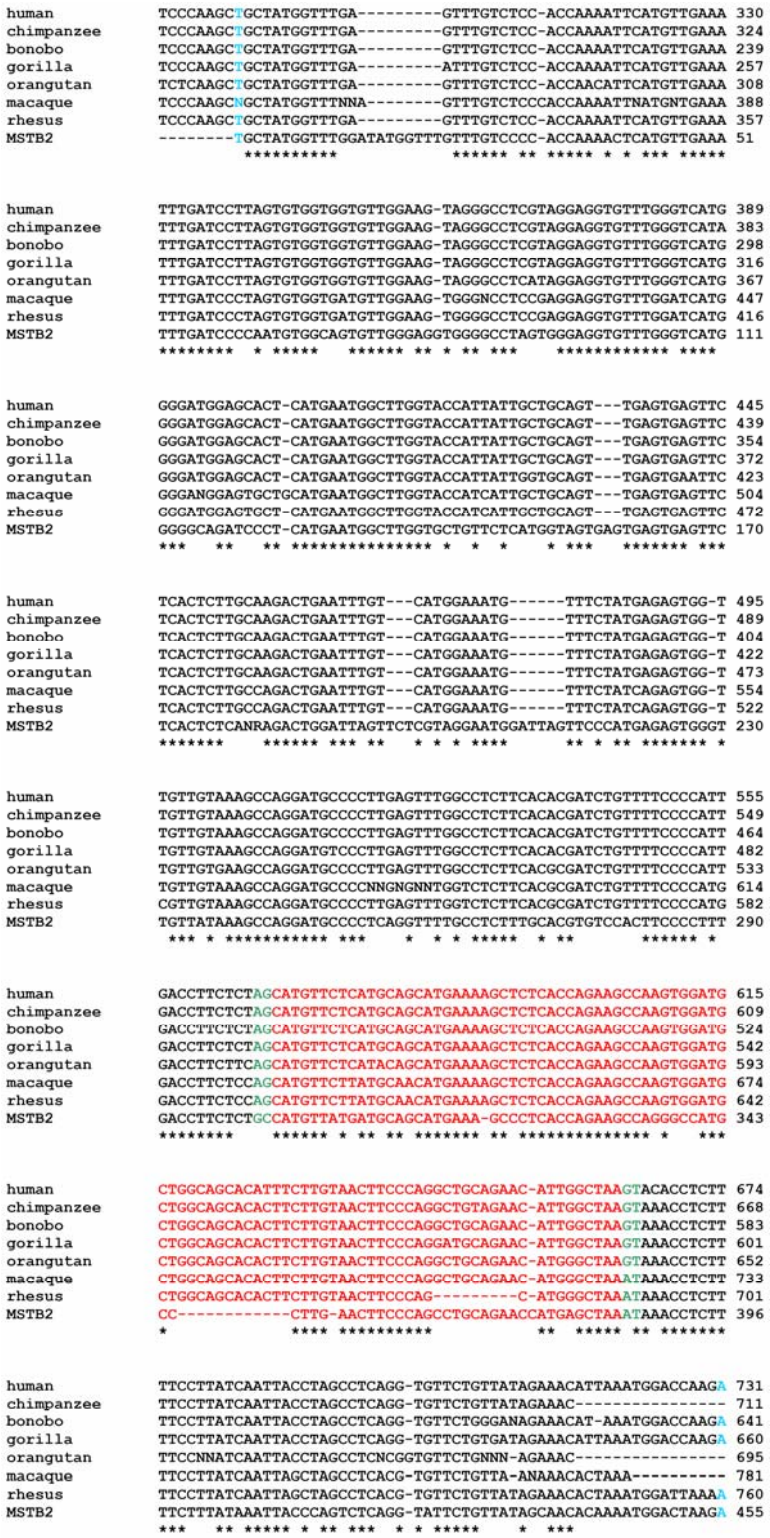

Figure 2.3: Multiple sequence alignment of PCR products. The PCR products are aligned and compared to the consensus sequence of MSTB2. The light blue letters indicate the start and end of LTR boundaries. The target exons and sequences in place of splice sites are shown in red and green color, respectively.

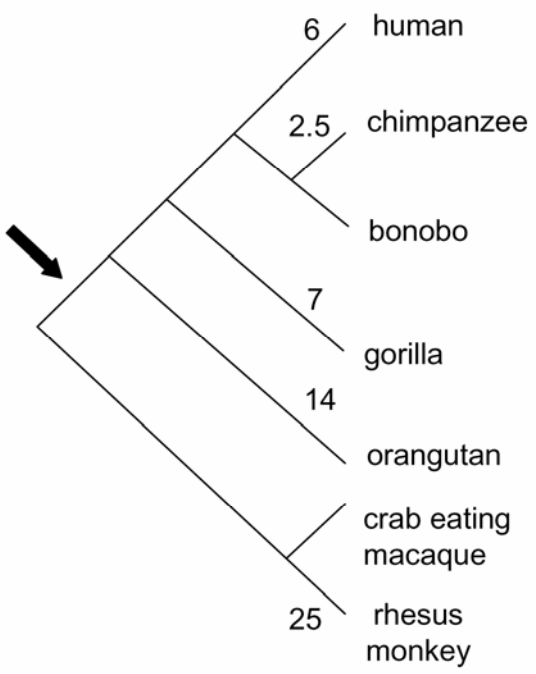

Figure 2.4: Evolutionary history of *IL22RA2* gene. A phylogenetic diagram of seven primates selected in this study. The numbers next to branches on the tree show the approximated divergence time from the last common ancestor in million-year time units (MYA). The arrow indicates the estimated point of emergence of the target exon caused by the LTR mutation from AT to GT making the canonical donor site.

To date, very little is known about the role and the origin of this additional exon (exon 3/4) in transcript variant 1. Being the only CRF2 protein with 32 amino acids inserted adjacent to the region important for ligand recognition, this isoform may bind to structurally different ligands than other isoforms (GRUENBERG *et al.* 2001). This possibility is supported by the experimental data which show that this variant fails to block IL22 activity (DUMOUTIER *et al.* 2001). The longer MaLR-related isoform may also modulate tissue-specific expression. The available data show that the *IL22RA2* isoform 1 is expressed only in placenta while isoform 2 is highly expressed in placenta and mammary gland and at a lower level in spleen, skin, thymus and stomach (GRUENBERG *et al.* 2001). However, nothing is known about the factors that control the expression of this

longest *IL22RA2* variant. Additional experiments should be performed to determine its function as well as to identify the possible change in ligand specificity due to the LTR-derived protein modification.

CONCLUSIONS

The distribution of LTR elements that became parts of human protein-coding genes shows the distinct preference of LRTS fixation in 5' and 3'untranslated regions. These observations confirm existing concept of LRTS role as a contributor to gene regulation evolution. On the other hand, the recruitment of LRTS to encode a part of a protein domain leading the exaptation to evolution of the host gene is a less frequent event. As shown in the part of this paper related to evolution of *IL22RA2* gene, several coincidences are necessary to allow the LRTS exonization event. The evolutionary analysis elucidates the action of the mechanism of incorporation of LRTS into a novel alternatively spliced exon.

METHODS

Bioinformatic analysis

The refGene file (hg17, May2004) with data on 18,241 RefSeq human genes (genes on chr_random excluded) including alternatively spliced variants (23,821 transcripts in total) was retrieved from the UCSC genome browser (KAROLCHIK *et al.* 2003). The annotations of 254,542 exons were compared with the transposable elements annotations available in the same database to determine the frequency of the LTR elements in the exon regions. The descriptions of the LTR elements were provided in the Repbase update (Repbase release 8.12) (JURKA 2000). We detected exon overlaps with

the LTR flanking regions and/or the internal sequences of LTR elements. The overlaps with exons were labeled as complete (LRTS covers the whole exon), partial (LRTS partially overlap with exon), or inside overlap (LRTS completely inside the exon). The type of exons associated with LRTS were then classified as the first CDS exon (first exon containing coding sequence), the last CDS exon (last exon containing coding sequence), single protein coding exon (exon containing the whole CDS of a gene), 5' UTR exon (exon located upstream to the first CDS exon/single protein coding exon) and 3' UTR exon (exon located downstream from the last CDS exon/single protein coding exon), and internal protein coding exon (all other CDS exons). In cases of the first and last CDS exons as well as single protein coding exons, LRTSs could be inserted in either the UTR and/or the CDS region. Finally, all the initial results were further processed. Exons identical in different transcripts were clustered to remove the redundancy. The LRTS fragments were reconstructed manually based on the initial data (*e.g.* LRTS family, human genome coordinates) and the LRTS information available in Repbase (JURKA 2000).

For all genes containing LRTS-derived exons, we used data of the Entrez gene and UCSC genome browser to infer information on alternative transcripts containing LRTS derived exons. Additionally, we checked the consistency of the reading frames in the exon overlaps with internal sequences of LRTS. The internal sequences of the LTR elements overlapping with CDS regions were translated in six reading frames and searched by BLAST and Pfam for the presence of domains of common viral proteins (*gag*, *pro*, *RT*, *RNaseH*, *IN* and *env*). The cases when detected viral protein domains became parts of human proteins were registered.

Given that the first and last CDS exons are commonly less reliably identified than the internal coding exons, we have considered further only 41 internal coding exons completely covered with LRTS. We have chosen exon 3/4 of the *IL22RA2* gene (splice variant 1) for in depth study using PCR-sequencing of homologous regions of several primate genomes and comparative analysis of the sequence data.

The primers flanking the target LTR-derived sequence were designed to be specific to the conserved human-chimpanzee-rhesus monkey region (data from the UCSC genome browser) using the PRIMER3 program (ROZEN and SKALETSKY 2000). Sequences of PCR fragments were aligned by the ClustalW program with default parameters (THOMPSON *et al.* 1994) and then were manually adjusted. For human, chimpanzee, and rhesus monkey, the annotated sequences of regions in question were previously available. In these three cases the known annotated sequences were used in the alignment while the PCR data were utilized as a complementary information. The donor-acceptor sites of the target exon were marked for all sequences based on the corresponding positions in the human *IL22RA2*. The timing of the exonization event was estimated via the phylogenetic analysis.

PCR amplification of the *IL22RA2* target exon

The PCR amplifications of genomic DNA of seven primate species (*Homo sapiens, Pan troglodytes, Pan paniscus, Gorilla gorilla, Pongo pygmaeus, Macaca fascicularis, Macaca mulatta*) were carried out by using the following primers, a forward primer 5'-ACCGCTACGACTTCTCTCTAC-3' and a reverse primer 3'-TCAGGTATTCTGGGGTCTG-5', which yield a 792 bp amplicon covering the region of the LTR in human. The PCR cycle conditions were as follows: initial 4 min and 30 sec

pre-denaturation at 94°C, 30 cycles of 30 sec denaturation at 94°C, 30 sec annealing at 50°C, 1 min elongation at 72°C, and a final 1-cycle extension of 7 min at 72°C. The PCR products were then purified on 1% (w/v) agarose gel, Gibco BRL Ultra-Pure, visualized by ethidium bromide staining and extracted by using the gel extraction kit (QIAGEN). Direct sequencing of the PCR products was performed by the DNA Sequencing Services, of the Genomics Core Facility at the Georgia Institute of Technology.

ACKNOWLEDGEMENTS

We would like to thank Alex Lomsadze for useful discussion of the computational procedure and King Jordan for helpful remarks on the final version of the manuscript. The work of JP and MB was supported in part by the NIH grant HG00783 to MB.

CHAPTER 3

EVALUATING THE PROTEIN CODING POTENTIAL OF EXONIZED TRANSPOSABLE ELEMENT SEQUENCES

ABSTRACT

Background

Transposable element (TE) sequences, once thought to be merely selfish or parasitic members of the genomic community, have been shown to contribute a wide variety of functional sequences to their host genomes. Analysis of complete genome sequences have turned up numerous cases where TE sequences have been incorporated as exons into mRNAs, and it is widely assumed that such 'exonized' TEs encode protein sequences. However, the extent to which TE-derived sequences actually encode proteins is unknown and a matter of some controversy. We have tried to address this outstanding issue from two perspectives: i-by evaluating ascertainment biases related to the search methods used to uncover TE-derived protein coding sequences (CDS) and ii-through a probabilistic codon-frequency based analysis of the protein coding potential of TE-derived exons.

Results

We compared the ability of three classes of sequence similarity search methods to detect TE-derived sequences among data sets of experimentally characterized proteins: 1-a profile-based hidden Markov model (HMM) approach, 2-BLAST methods and 3-RepeatMasker. Profile based methods are more sensitive and more selective than the

other methods evaluated. However, the application of profile based search methods to the detection of TE-derived sequences among well-curated experimentally characterized protein data sets did not turn up many more cases than had been previously detected and nowhere near as many cases as recent genome-wide searches have. We observed that the different search methods used were complementary in the sense that they yielded largely non-overlapping sets of hits and differed in their ability to recover known cases of TE-derived CDS. The probabilistic analysis of TE-derived exon sequences indicates that these sequences have low protein coding potential on average. In particular, non-autonomous TEs that do not encode protein sequences, such as Alu elements, are frequently exonized but unlikely to encode protein sequences.

Conclusions

The exaptation of the numerous TE sequences found in exons as *bona fide* protein coding sequences may prove to be far less common than has been suggested by the analysis of complete genomes. We hypothesize that many exonized TE sequences actually function as post-transcriptional regulators of gene expression, rather than coding sequences, which may act through a variety of double stranded RNA related regulatory pathways. Indeed, their relatively high copy numbers and similarity to sequences dispersed throughout the genome suggests that exonized TE sequences could serve as master regulators with a wide scope of regulatory influence.

BACKGROUND

Transposable elements (TEs) are DNA sequences capable of moving (transposing) among locations in the genomes of their host organisms. When TEs transpose they often replicate themselves and they can accumulate to very high copy

numbers. For instance, at least 47% of the human genome is made up of TE-derived sequences (LANDER *et al.* 2001). For many years, TEs were thought to be genomic parasites that did not contribute functionally relevant sequences to the genomes in which they reside (DOOLITTLE and SAPIENZA 1980; ORGEL and CRICK 1980). However, as of late it has become increasingly apparent that TEs can have profound effects on the structure, function and evolution of their host genomes (BIEMONT and VIEIRA 2006; JURKA *et al.* 2007; KAZAZIAN 2004; KIDWELL and LISCH 2001).

One way that TEs have contributed to the function and evolution of their host genomes is through the donation of regulatory sequences that control the expression of nearby genes. This phenomenon was originally noticed through the elucidation of individual cases where host genes were found to be regulated by TE-derived sequences (BRITTEN 1996; BRITTEN 1997). Later, genome-scale analyses confirmed that TE-derived sequences have contributed diverse and abundant regulatory sequences to host genomes (JORDAN *et al.* 2003; VAN DE LAGEMAAT *et al.* 2003).

TEs can also contribute to host genomes by providing protein coding sequences. This process is initiated when a new or existing TE sequence becomes captured as an exon (exonized) in a host gene mRNA sequence. The exonization of TE sequences appears to be quite common in eukaryotic genomes. An early highthroughput analysis of the human transcriptome by Nekrutenko and Li revealed that 4% of human protein coding regions contained TE sequences (NEKRUTENKO and LI 2001). However, the extent to which exonized TE sequences actually contribute *bona fide* protein coding sequences has been called into question. It is simply not clear whether the presence of a TE

sequence in a spliced exon, *i.e.* as part of an mRNA, indicates that it will ultimately be translated into a functioning protein.

Two reports in particular have challenged the figure of 4% of human proteins with TE-derived coding sequences. In both of these studies, more conservative approaches to the identification of TE-derived protein coding sequences were taken. Specifically, these studies employed the analysis of coding sequences taken exclusively from proteins that had been experimentally characterized, either through elucidation of their 3D structures or via direct peptide sequencing methods. Thus, only the best characterized protein coding sequences were studied and gene predictions, or models, based on the mapping of expressed sequences to genomes were not considered. This approach was first taken by Pavlicek *et al.* who surveyed a dataset of 781 non-redundant human proteins with 3D structures for the presence of TE-derived coding sequences (PAVLICEK *et al.* 2002). They were not able to find a single reliable case of a TE-derived protein coding sequence in these data. Considering these results together with the previous work of Nekrutenko and Li (NEKRUTENKO and LI 2001), the authors concluded that while many alternative transcripts may include TE sequences, these are rarely if ever incorporated into the mRNA sequences that are destined to be translated into proteins. Pavlicek *et al.* found it particularly unlikely that non-coding TEs, such as Alu elements, could evolve to encode proteins after being incorporated into host mRNAs.

Gotea and Makalowski conducted a similar, if further reaching, study by looking for TE-derived sequences in the coding regions of human proteins taken from the Protein Data Bank (BERMAN *et al.* 2000) (3,764) and from the Swiss-Prot (BOECKMANN *et al.* 2003) collection of directly sequenced human peptides (1,765) (GOTEA and

MAKALOWSKI 2006). Evaluation of these sequences with the RepeatMasker program (SMIT *et al.* 1996-2004) uncovered 24 cases of TE-derived protein coding sequences. However, many of these had relatively low sequence similarity scores that were close the RepeatMasker threshold for false-positives. After further evaluation of these cases using a variety of comparative sequence analysis techniques, the authors settled on a figure of 0.1% for the percentage of actual protein coding sequences with TE-derived exons. Incidentally, this figure is in line with the initial analysis of the human genome sequence, which found 47 cases of human protein coding regions with TE-derived sequences, corresponding to ~0.16% of all human genes given the total human gene number count of ~30,000 used at that time (LANDER *et al.* 2001).

While there can be little doubt that these two aforementioned studies used appropriately conservative datasets to search for TE-derived protein coding sequences, it may also be the case that the primary detection methods they employed are insufficiently sensitive since they rely on DNA-DNA sequence comparisons. For instance, RepeatMasker, which is the most widely used program for the detection of TE sequences, uses pairwise comparisons of genomic DNA sequences with DNA consensus sequences that represent TE families. Protein sequence based similarity searches are more sensitive than DNA based searches, and profile searches that take advantage of information on site-specific variation along protein domains are proven to be the most sensitive approach for detecting sequence homology (EDDY 1996; EDDY 1998; SONNHAMMER *et al.* 1997).

The increased sensitivity of protein and profile based searches is underscored by two recent studies that uncovered many more putative cases of TE-derived protein

coding sequences. Roy Britten compared human protein coding sequences to the Repbase library of consensus TE sequences (JURKA 2000; JURKA *et al.* 2005) using both RepeatMasker and a protein sequence based approach that used six-frame translations of Repbase sequences. Use of the protein (translated) sequence search method resulted in a more than two-fold increase, from 814 to 1,950, in the number of genes found to have TE-derived protein coding sequences (BRITTEN 2006). An even more sensitive profile based search method was used by Zdobnov *et al.* to search for TE-derived protein coding sequences in four vertebrate genomes (ZDOBNOV *et al.* 2005). These authors compiled a set of known protein domains that are characteristic of TEs, and profiles of these domains were then used in hidden Markov model (HMM) searches of the protein sequences. This analysis resulted in the discovery of 1,000 vertebrate genes containing protein coding sequences that are related to TEs. However, neither the Britten nor the Zdobnov *et al.* studies confined their searches to experimentally characterized protein coding sequences as did the studies of Pavlicek *et al.* and Gotea and Makalowski, both of which resulted in far smaller estimates for the fraction of genes with TE-derived protein coding sequences.

Clearly, the extent to which TEs contribute protein coding sequences to vertebrate genomes is not a settled matter. Relatively insensitive searches of conservative data sets lead to low estimates for the fraction of TE-derived protein coding sequences, while more sensitive searches of less conservative data sets yield higher fractions. The aims of this study are i-to evaluate the ascertainment biases related to different sequence similarity search methods and ii-to try and better understand the potential of TEs to contribute protein coding sequences to vertebrate genomes. To these ends, we searched conservative, experimentally characterized, protein coding sequence data sets for TE-

derived sequences using sensitive profile based search methods. We also compared the results of profile based search methods with more traditional pairwise DNA and protein based search methods. Known cases of experimentally characterized proteins with TE-related sequences were used as positive controls to assess the sensitivity of the different sequence similarity search techniques. Finally, we used probabilistic gene prediction methods as well as an analysis of relative nucleotide (GC) frequencies across codon positions to evaluate the protein coding probability of TE-derived exon sequences.

RESULTS AND DISCUSSION

Searching for TE-associated proteins

We used a number of approaches to detect molecular domestication events, specifically exaptation of host (cellular) CDS from TE sequences, by searching for the presence of TE-related sequences in functionally well characterized host protein sequences and CDS. A total of 41,492 PDB entries and 21,050 Swiss-Prot directly sequenced proteins were taken to represent functionally well characterized proteins (genes) since they have been experimentally determined. Viral proteins were excluded from these data sets in order to avoid the overlap among protein domains shared between viral and retrotransposon-encoded proteins resulting in final data sets of 39,252 PDB and 20,732 Swiss-Prot entries. Using the combined automatic and manual search procedure described in the Methods section, we identified 124 TE related Pfam protein domains (See Table B.1). We then searched for the presence of these TE-related domains among the experimentally characterized PDB and Swiss-Prot data sets using profile-based similarity search methods (HMM profiles) as described in the Methods section. The numbers (percentages) of protein sequences found to possess TE-related domains, based

on a series of increasingly stringent HMM search cutoff criteria, are shown in Table 3.1 and Table 3.2 for the PDB and Swiss-Prot data sets respectively.

Table 3.1: Detection of TE-encoded sequences in PDB proteins
The number of PDB entries found with TE protein fragments (from autonomous TEs) by different search programs is shown. The percentage of total PDB entries is shown in the parenthesis. The square bracket indicates the number and the percentage of protein entries associated with sequences derived from TEs including the non-autonomous ones.

Cutoff value	HMMER	BLASTN	BLASTP	BLASTX	TBLASTN	TBLASTX	RM
E-value ≤ 1	17543 (44.69%)	4924 (16.15%) [10890: 35.72%]	1558 (3.97%)	1643 (4.19%) [3343: 8.52%]	8662 (28.41%)	13107 (42.99%) [19891: 65.25%]	N/A
E-value ≤ 0.1	2757 (7.02%)	1531 (5.02%) [3076: 10.09%]	614 (1.56%)	764 (1.95%) [1207: 3.08%]	5671 (18.60%)	7721 (25.33%) [11474: 37.64%]	N/A
E-value ≤ 0.01	533 (1.36%)	778 (2.55%) [1453: 4.77%]	424 (1.08%)	586 (1.49%) [827: 2.11%]	3943 (12.93%)	5688 (18.66%) [8481: 27.82%]	N/A
E-value ≤ 0.001	256 (0.65%)	564 (1.85%) [917: 3.01%]	364 (0.93%)	530 (1.35%) [700: 1.78%]	3030 (9.94%)	4832 (15.85%) [6995: 22.94%]	N/A
E-value ≤ 0.0001	168 (0.43%)	423 (1.39%) [682: 2.24%]	308 (0.78%)	464 (1.18%) [552: 1.41%]	2266 (7.43%)	4057 (13.31%) [6033: 19.79%]	N/A
E-value ≤ 0.00001	148 (0.38%)	371 (1.22%) [555: 1.82%]	210 (0.54%)	388 (0.99%) [474: 1.21%]	1676 (5.50%)	3533 (11.59%) [5035: 16.52%]	N/A
GA (gathering threshold)	140 (0.36%)	N/A	N/A	N/A	N/A	N/A	N/A
TC (trusted cut offs)	140 (0.36%)	N/A	N/A	N/A	N/A	N/A	N/A
default value	N/A	N/A	N/A	N/A	N/A	N/A	465 (1.53%) [950: 3.12%]

Table 3.2: Detection of TE-encoded sequences in Swiss-Prot directly sequenced proteins

The number of Swiss-Prot directly sequenced proteins found with TE protein fragments (from autonomous TEs) by different search programs is shown. The percentage of total Swiss-Prot entries is shown in the parenthesis. The square bracket indicates the number and the percentage of protein entries associated with sequences derived from TEs including the non-autonomous ones.

Cutoff value	HMMER	BLASTN	BLASTP	BLASTX	TBLASTN	TBLASTX	RM
E-value ≤ 1	8182 (39.47%)	2108 (16.39%) [4468: 34.74%]	2909 (14.03%)	3030 (14.62%) [4620: 22.28%]	3418 (26.58%)	5052 (39.28%) [7576: 58.91%]	N/A
E-value ≤ 0.1	1368 (6.60%)	655 (5.09%) [1331: 10.35%]	1481 (7.14%)	1632 (7.87%) [2185: 10.54%]	2159 (16.79%)	3009 (23.40%) [4501: 35.00%]	N/A
E-value ≤ 0.01	214 (1.03%)	316 (2.46%) [573: 4.46%]	935 (4.51%)	1103 (5.32%) [1503: 7.25%]	1559 (12.12%)	2180 (16.95%) [3208: 24.95%]	N/A
E-value ≤ 0.001	65 (0.31%)	187 (1.45%) [351: 2.73%]	694 (3.35%)	844 (4.07%) [1186: 5.72%]	1204 (9.36%)	1863 (14.49%) [2668: 20.75%]	N/A
E-value ≤ 0.0001	30 (0.14%)	112 (0.87%) [235: 1.83%]	516 (2.49%)	668 (3.22%) [971: 4.68%]	882 (6.86%)	1607 (12.50%) [2236: 17.39%]	N/A
E-value ≤ 0.00001	19 (0.09%)	83 (0.65%) [181: 1.41%]	372 (1.79%)	516 (2.49%) [776: 3.74%]	653 (5.08%)	1403 (10.91%) [1926: 14.98%]	N/A
GA (gathering threshold)	14 (0.07%)	N/A	N/A	N/A	N/A	N/A	N/A
TC (trusted cut offs)	14 (0.07%)	N/A	N/A	N/A	N/A	N/A	N/A
default value	N/A	N/A	N/A	N/A	N/A	N/A	154 (1.20%) [336: 2.61%]

To compare the sensitivity of the HMM profile-based search method with more standard sequence-against-sequence similarity search methods, we used the BLAST and RepeatMasker programs to search for TE-derived sequences among host proteins and their corresponding CDS. To this end, we built CDS databases corresponding to the PDB

and Swiss-Prot protein data sets, which contain 34,795 and 38,754 CDS sequences, respectively. These CDS data sets correspond to 30,486 PDB and 12,860 Swiss-Prot proteins. The difference in the number of proteins versus CDS can be attributed to the fact that a number of protein sequences lack the matching CDS because they are synthetic, mutated, or chimeric proteins. In addition, some protein entries may be related to more than one CDS sequence, while some CDS may match with several PDB entries due to the redundancy of protein chains. For use as query sequences in BLAST searches, we created three TE sequence libraries from data provided in Repbase: 5,611 TE sequences (for all TEs in all taxa), 1,423 TE encoded proteins and 1,349 TE CDS sequences. The specific combinations of BLAST program, query set and data base set used in each search is shown in Table 3.3. The numbers (percentages) of sequences found with TE-related domains, based on a series of increasingly stringent E-value cutoffs, are shown in Table 3.1 and Table 3.2 for the PDB and Swiss-Prot data sets respectively. Finally, the RepeatMasker program was used to search for TE-related sequences among the PDB and Swiss-Prot CDS data sets (see numbers and percentages of hits in Table 3.1 and Table 3.2).

Considering the results of the three different classes of search strategies – HMMER, BLAST and RepeatMasker – together yields some unexpected results. Not surprisingly, however, RepeatMasker proved to be the least sensitive strategy to search for TE-related host protein coding sequences. Using the fairly liberal default cutoff value, which returns a number of hits with marginal reliability, RepeatMasker yields a lower number of hits than all but the most conservative searches with the other methods (Table

3.1 and Table 3.2). This is consistent with the fact that RepeatMasker relies on DNA-DNA sequence comparison.

Table 3.3: Sequence similarity program-query-database combinations used to search for TE-related host sequences

Tool	Query	Database
HMMER	PDB/Swiss-Prot protein	HMM profiles of TE-related Pfam domains
BLASTN	TE CDS & all TE sequences	PDB/Swiss-Prot CDS
BLASTP	TE protein	PDB/Swiss-Prot protein
BLASTX	TE CDS & all TE sequence	PDB/Swiss-Prot protein
TBLASTN	TE protein	PDB/Swiss-Prot CDS
TBLASTX	TE CDS & all TE sequence	PDB/Swiss-Prot CDS
RepeatMasker	PDB/Swiss-Prot CDS	TE CDS & all TE sequences

To compare the results of the HMMER versus BLAST search strategies, we plotted the percentage of hits against the E-value threshold used (Figure 3.1A and 3.1B). Together with Tables 3.1 and 3.2, these plots show the relative numbers (percentages) of hits retrieved using each method. TBLASTX searches, where CDS are translated in all six reading frames and are searched against nucleotide databases that are translated in six frames, gave the highest number of hits across all but the most liberal E-value cutoffs. This is consistent with previous results, showing that translated BLAST searches yield far more TE-host protein similarity than BLASTN or RepeatMasker searches (BRITTEN 2006). The profile-based HMMER searches, which are expected to be the most sensitive, did return the highest number of hits at liberal E-values, but after two rounds of decreasing E-values, HMMER dropped off to yield the fewest number of hits across all the methods (Table 3.1, Table 3.2 and Figure 3.1). Thus, HMMER appears to be particularly sensitive to increasingly stringent E-value cutoffs.

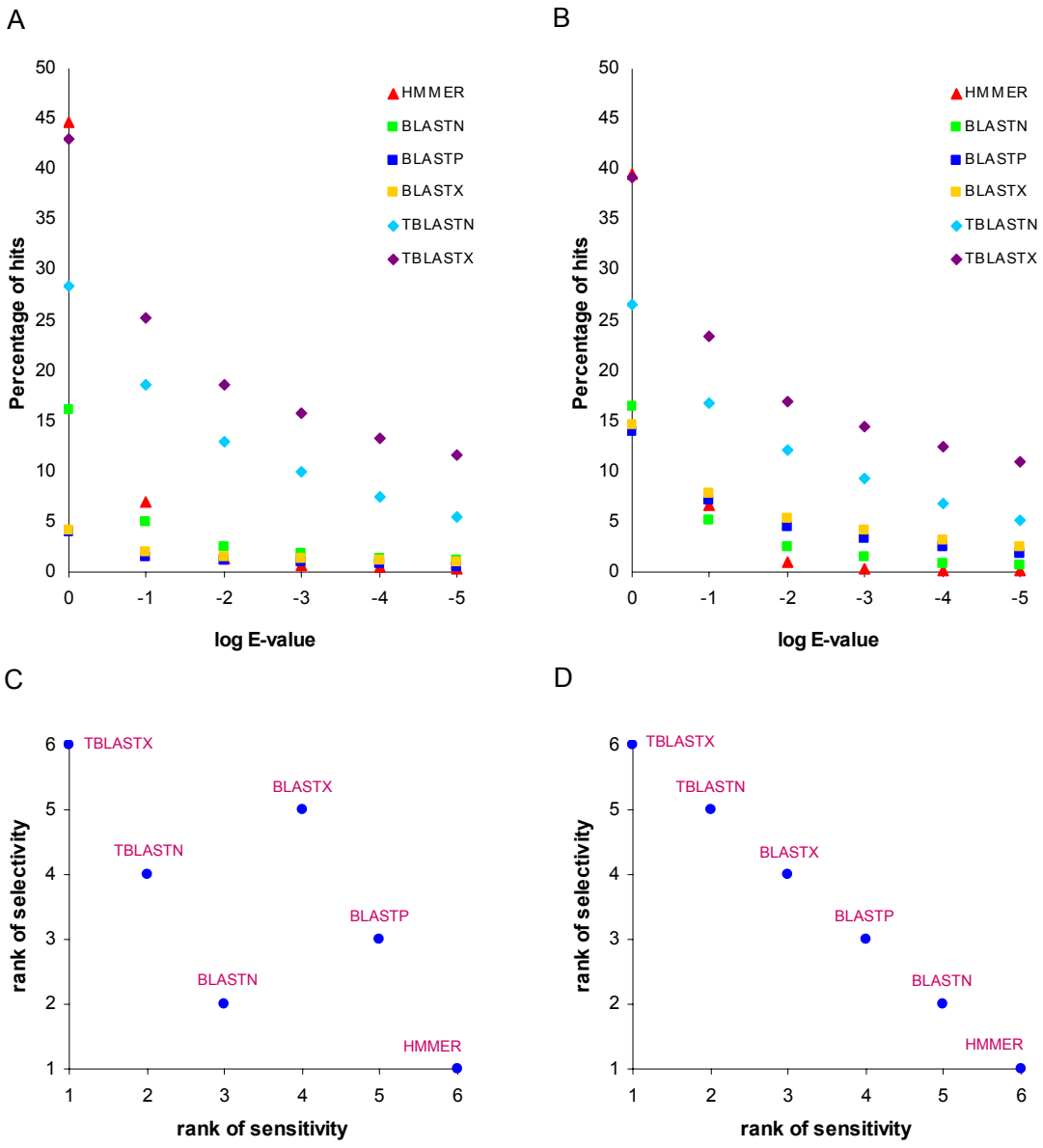

Figure 3.1: Sensitivity and selectivity comparison for different sequence similarity search methods. The percentage of hits returned by different sequence similarity search methods are shown across increasingly stringent E-value cutoffs for the PDB (A) and Swiss-Prot (B) data sets. The selectivity and sensitivity ranks are compared for different search methods for the PDB (C) and Swiss-Prot (D) data sets.

To evaluate the selectivity of the search methods we employed, we measured the exponential rate of decline in the relative number (percentage) of hits retrieved at

decreasing E-value thresholds, which allowed us to measure the effect of increasing stringency on the number of hits retrieved across methods. This was done by fitting exponential trend lines to the data shown in Figure 3.1A and Figure 3.1B and then ranking the searches with respect to the exponent of the trend line; the most selective methods are ranked the highest (*i.e.* have the lowest rank number). In this way, HMMER was shown to be the most selective method and TBLASTX the least selective. As could be expected, selectivity is inversely correlated with sensitivity, and exactly so for the Swiss-Prot search, as can be seen when the ranks of method sensitivity (number of hits) are compared to the selectivity ranks (Figure 3.1C and Figure 3.1D). Again, this overall trend defied the expectations of increased sensitivity of profile methods that we had at the outset of the study.

We also considered the relationships among the different search methods in terms of the fraction of hits that they had in common. For each pair of search methods, the fraction of shared hits was calculated (see Methods), and the resulting pairwise similarity matrix was used to cluster the methods (Figure 3.2). For both the PDB and Swiss-Prot searches, the translated BLAST methods group together as do the protein searches BLASTP and BLASTX. BLASTN was more similar to the translated methods in the PDB search, while it had lower overlap with the other BLAST methods in the Swiss-Prot search. HMMER consistently showed the lowest overlap with other methods. Perhaps more importantly, the extent of overlap between the different methods was surprisingly low. For instance, at the lowest E-value cutoff only 2 out of a total of 4,241 hits for PDB and 2 out of 1,724 for Swiss-Prot were identified by all six search methods. This underscores the fact that the different search methods are very much complementary and

indicates that an exhaustive search for potential TE-CDS exaptation events will require the use of a variety of search techniques.

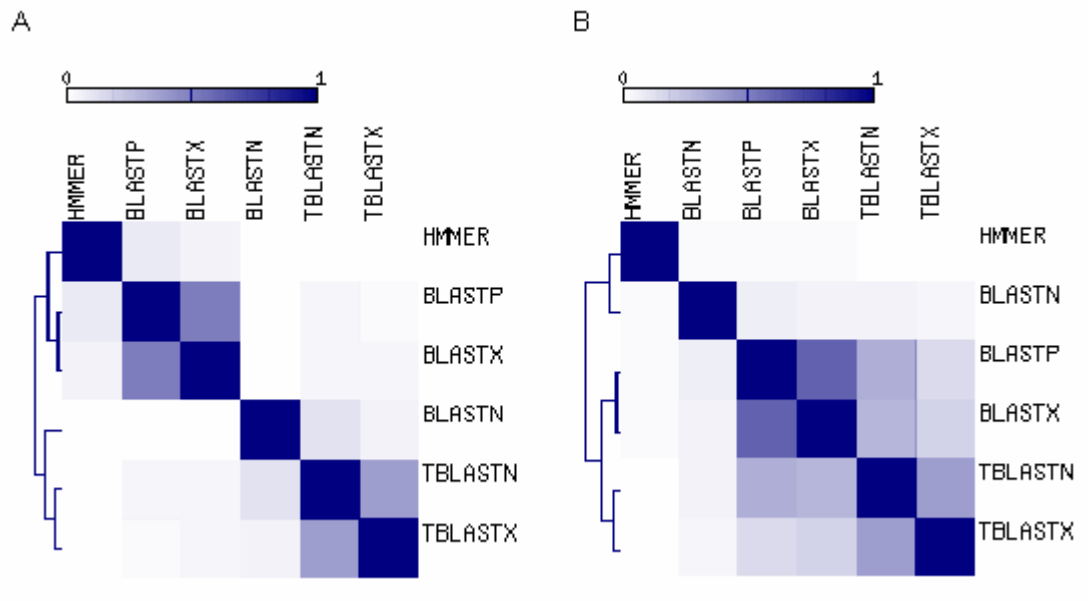

Figure 3.2: Relationships among sequence similarity search methods. Colors represent the fraction hits shared between methods, from 0 (white) to 1 (purple). The matrices are symmetrical with self-similarity shown along the diagonal. The search methods are ordered along both axes of the plots with respect to similarity, and dendograms showing the relationships among methods are shown for the PDB (A) and Swiss-Prot (B) data sets.

Comparative analysis of cases of TE-CDS exaptation

HMMER was also run using the most conservative gathering (GA) and trusted cutoff (TC) thresholds described in the Methods section. Searches using GA and TC yield the fewest number of hits for both the PDB and Swiss-Prot searches. Thus, we took these results to be the most reliable (conservative) set of TE-related host proteins and further evaluated these results to look for *bona fide* cases of TE-CDS exaptation.

By manually evaluating these results, we were able to classify the hits into five

distinct categories (see Methods), only one of which represents the kinds of TE-CDS exaptation events that we are most interested in (Table 3.4). For instance, the vast majority of apparent TE-related proteins in the PDB data set corresponded to either synthetic constructs (*i.e.* artificial sequences) or non-specific, and often ubiquitous, TE-related protein domains such as RNase H. For this latter category, the non-specific TE-related domains, it is a formal possibility that they represent ancient TE-CDS exaptation events but it is difficult, if not impossible, to unambiguously support that assertion. Other proteins detected in the PDB set correspond to TE-encoded proteins and viral proteins. Only 11 out of 140 cases (or 7.9%) correspond to likely TE-CDS exaptation events. With the GA and TC thresholds, the Swiss-Prot dataset yielded far fewer total hits than did PDB and only 3 of these correspond to likely TE-CDS exaptation events (Table 3.4).

Table 3.4: Classification of proteins containing TE-associated Pfam domains detected by the GA and TC cutoffs of HMMER
The categories of hits are described in the text and the number (percentage) for each category is shown for searches against the PDB and Swiss-Prot data sets.

Category	PDB	Swiss-Prot
Potential TE-related proteins	11 (7.86%)	3 (21.43%)
Viral proteins	14 (10.00%)	0 (0%)
TE-encoded proteins	18 (12.86%)	7 (50.00%)
Synthetic construct	47 (33.57%)	0 (0%)
non-specific TE-related protein domains	50 (35.71%)	4 (28.57%)

A set of 12 likely TE-CDS exaptation events, representing the non-redundant union of the most reliable cases from the PDB and Swiss-Prot sets in Table 3.4, were further analyzed in order to assess the ability of BLAST and RepeatMasker to detect these cases. Only one of the 12 proteins was detected using all methods, and again,

RepeatMasker was shown to be the least sensitive method (Table 3.5). Indeed, as expected, DNA-DNA search methods in general were found to be insensitive; there are 4 cases where BLASTN and RepeatMasker are the only programs unable to detect the TE-CDS similarity. There were four individual cases, corresponding to two different Pfam domains, where only HMMER was able to detect the TE-protein sequence similarity. These results stand in contrast to the results of the previous section, which indicate that HMMER is the least sensitive search method overall. There are two possible explanations for this dissonance. First of all, HMMER may suffer from a lack of coverage due to its reliance on the collection of Pfam domain family definitions. Secondly, and perhaps more plausible, the different search methods may in fact be complementary in terms of detecting different sets of exaptation events. This may be particularly relevant for DNA based, and/or translated, search methods that are able to compare non-coding TE-derived sequences to host protein and CDS sequences.

Table 3.5: Analysis of the qualified set of TE-associated domain containing proteins
Twelve PDB/Swiss-Prot proteins with TE-associated Pfam domains detected by HMMER (GA and TC cutoffs) are shown. The results from BLAST and RepeatMasker analysis are compared (√ = found, X = not found TE-related sequence). The cutoff E-value of 0.01 was used as the detection criteria.

Accession	Name	Organism	Pfam domain	BLASTN	BLASTP	BLASTX	TBLASTN	TBLASTX	RM
2jm3	Hypothetical protein	*C. elegans*	THAP	X	X	X	X	X	X
1a0p, XERD_ECOLI	Tyrosine recombinase xerD	*E. coli*	Phage_integrase	X	X	X	X	X	X
1bw6, 1hlv	Centromere protein B	*H. sapiens*	CENP-B_N	√	√	√	√	√	X
1uhu	retroviral Gag MA-like domain of RIKEN cDNA 3110009E22	*M. musculus*	Gag_MA	N/A	√	N/A	√	N/A	N/A
1y4m	Syncytin-2	*H. sapiens*	TLV_coat	√	√	√	√	√	√
2a3v	Site-specific recombinase IntI4	*V.cholerae*	Phage_integrase	X	X	X	X	X	X
2cqf	Lin-28 homolog A (Zinc finger CCHC domain-containing protein 1)	*H. sapiens*	zf-CCHC	X	√	√	√	√	X
2ct5	Zinc finger BED domain-containing protein 1	*H. sapiens*	zf-BED	X	√	√	√	√	X
2d8r	THAP domain-containing protein 2	*H. sapiens*	THAP	X	√	√	√	√	X
2djr	Zinc finger BED domain-containing protein 2	*H. sapiens*	zf-BED	X	√	√	√	√	X
CBH1_SCHPO	CENP-B homolog protein 1	*S. pombe*	DDE	X	√	√	√	√	X
XERC_ECOLI	Tyrosine recombinase xerC	*E. coli*	Phage_integrase	X	X	X	X	X	X

Case studies of known TE-derived genes

There are a number of well verified cases of host proteins (genes) that are known to have been derived from TE sequences. These are proteins that have been shown to be functionally analogous and evolutionarily derived from their TE-encoded counterparts. For instance, the enzyme Telomerase evolved from TE-encoded reverse transcriptase enzymes (BLACKBURN 1991; EICKBUSH 1997) and the RAG1 recombinase is related to the transposase enzymes (AGRAWAL *et al.* 1998; KAPITONOV and JURKA 2005). The centromere protein CENP-B (KIPLING and WARBURTON 1997) and SETMAR (CORDAUX *et al.* 2006) are other well documented cases of the evolution of host CDS from TEs. We have used these cases as positive controls in order to further evaluate the ability of the different classes of search methods to detect cases of TE-CDS exaptation. We assessed the ability of each program to detect human proteins or CDS for all four of these cases (Table 3.6). Translated BLAST searches BLASTX and TBLASTN were the most sensitive search methods finding all of the cases in this data set, and HMMER was shown to be fairly sensitive in detecting three out of four of the known cases of TE-exaptation. RepeatMasker was the least sensitive detecting only the SETMAR case. SETMAR represents an evolutionarily recent TE-CDS exaptation event that occurred during the primate radiation some 40-58 million years ago (CORDAUX *et al.* 2006). Thus, the SETMAR CDS retains DNA sequence similarity to the Hsmar1-type TE transposase gene from which it is derived. In any case, all the search methods were able to detect SETMAR, so RepeatMasker would not be necessary to elucidate this case. In general, for the BLAST searches, translated and protein based searches are the most sensitive followed by DNA-based BLASTN.

Table 3.6: Detection of previously identified TE-associated proteins
The ability of the different sequence similarity search methods to detect well known cases of TE-derived CDS is indicated with √ and failure to detect is indicated with X.

Name	TE-protein	HMMER	BLASTN	BLASTP	BLASTX	TBLASTN	TBLASTX	RM
Telomerase	Reverse transcriptase (LINEs)	√	X	X	√	√	√	X
RAG1	Transposase (Transib superfamily)	X	X	√	√	√	X	X
CENP-B	pogo-like DNA transposase	√	√	√	√	√	√	X
SETMAR	Hsmar1 transposase	√	√	√	√	√	√	√

Evolutionary relationship between TE and cellular proteins

In the formal sense, establishing a solid, statistically significant, sequence similarity relationship between TE-encoded and cellular proteins is necessary but not sufficient to make the claim of a TE-CDS exaptation event. This is exemplified by the numerous cases of ubiquitous, non-specific TE-related protein domains uncovered when searching the PDB and Swiss-Prot experimentally characterized data sets (Table 3.4). These abundant protein domains, such as RNase H, can be functional analogs that have evolved convergently in host and TE genomes or they may have their evolutionary origins in host (cellular) genomes and been subsequently captured by TEs. Thus, it is necessary to document the evolutionary relationships between TE encoded and related host-encoded protein domains as accurately as possible in order to evaluate the evidence for TE-CDS exaptation. Phylogenetic analysis is best suited to this task. Indeed, phylogenetic analysis is needed to unequivocally demonstrate a TE-origin, *i.e.* the direction of the TE-to-host sequence transfer, for protein domains with similarity between TEs and host genomes as was shown for the case of Telomerase (EICKBUSH 1997).

To illustrate this analytical process, we have chosen the THAP protein domain. Sequence similarity between the THAP domain and TEs has been noted previously but the evolutionary origins of the domain, and in particular the specific direction of the TE-host transfer, remains uncertain. The *Caenorhabditis elegans* C terminal binding protein (CtBP) [PDB: 2jm3] contains the THAP domain, a ~90 residue domain, which is restricted to animals and shared between the THAP family of cellular DNA-binding proteins and transposases encoded by DNA-type TEs. This domain was previously found to be homologous to the site-specific DNA-binding domain (DBD) of *Drosophila* P-element transposase (ROUSSIGNE *et al.* 2003). An evolutionary analysis of the domain architectures and sequence similarities among THAP domain containing proteins was taken to suggest the possibility that cellular proteins have recruited this domain on more than one occasion (QUESNEVILLE *et al.* 2005). In order to characterize all sequence relationships between TE and host encoded THAP domains, we used HMMER with the Pfam THAP domain HMM profile to search among the Repbase library of TE-encoded proteins. The use of HMMER was necessitated by the fact that, consistent with results reported in previous sections, BLAST and RepeatMasker can not detect any TE-related sequence in *C. elegans* CtBP. Using HMMER, we found that PROTOP is the identity of the autonomous *Drosophila melanogaster* P element that contains the THAP domain, in positions 12 to 94 of its consensus protein sequence. We also identified six additional TE families containing THAP domain (KBOC_DB, P1_AG, P3_AG, P4_AG, Kolobok-1_XT, Kolobok-2_BF). In addition, CtBP was used as a BLASTP search query to identify host (cellular) genome encoded THAP domains. All TE and host encoded THAP

domains were aligned, globally and locally, and phylogenetically analyzed as described in the Methods section.

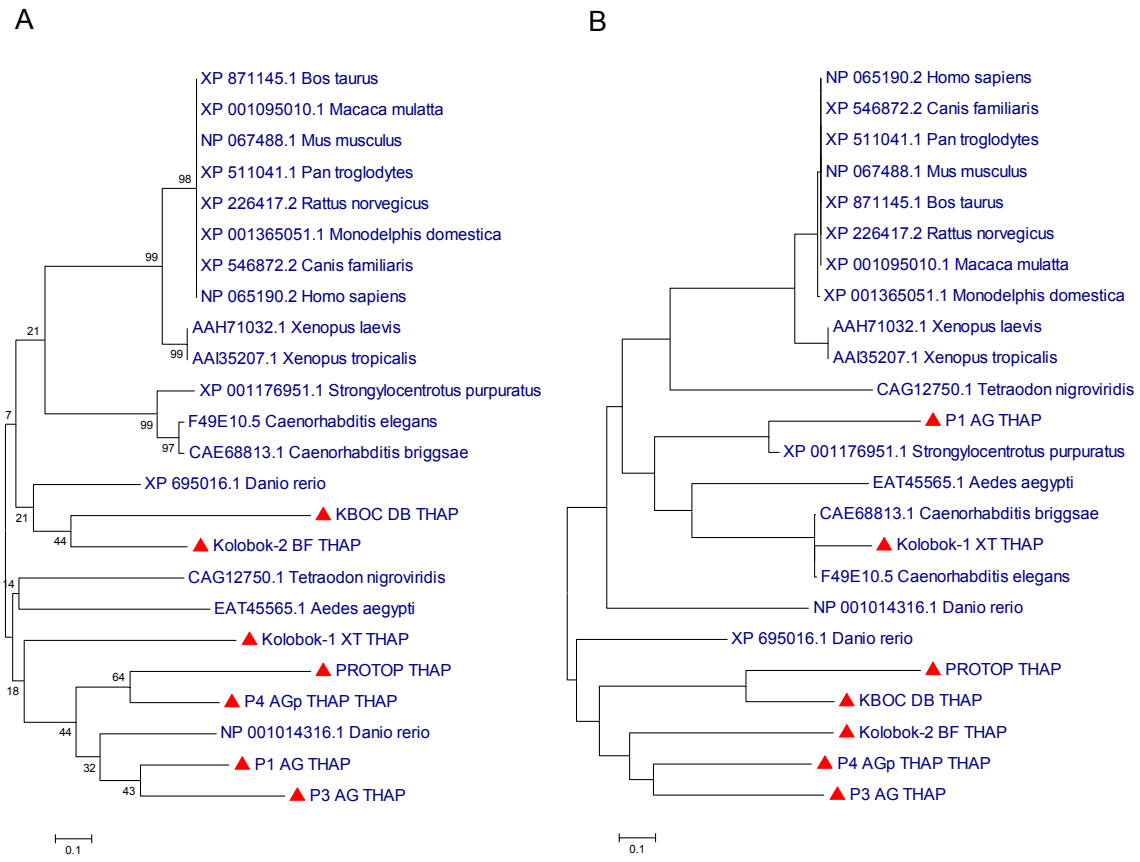

Figure 3.3: Phylogenetic relationship of TE and cellular THAP domains. Neighbor-joining trees of seven THAP homologous TE sequences and seventeen cellular THAP sequences from various species are shown. The trees were created based on (A) the multiple sequence alignment of all THAP sequences and (B) the pairwise gamma distance matrix calculated from BLAST all-against-all pairwise alignments. TE-THAP sequences are indicated by red triangle marks. Bootstrap values (A) represent the fraction of times that internal branches, supporting clades, were recovered among trees built from 1,000 re-sampled alignments.

The global and local alignment based phylogenetic analyses consistently identify one clade of host encoded THAP domains and a second clade of THAP domains encoded by both TEs and cellular genomes (Figure 3.3). Interestingly, the TE and host encoded

domains are distributed throughout this clade suggesting the possibility of multiple transfers of THAP domain CDS between TE and host genomes. In addition, TE encoded THAP domains appear to have greater sequence diversity, reflected by the branch lengths, than host encoded THAP domains, consistent with a TE origin of the domain. Thus, it appears that THAP indeed evolved among TE sequences and was subsequently transferred on more than one occasion to host (cellular) genomes.

Protein coding potential of TE-derived exons

By now, it is well known that TE-derived sequences are frequently incorporated into the exons of host mRNAs (NEKRUTENKO and LI 2001). What is less clear is the extent to which TE-derived exons of host genes are destined to become protein coding sequences. Previously, we addressed this issue by searching functionally well characterized protein coding sequences for the presence of TE-related domains. Here, we take a DNA sequence codon based approach to this question. Our approach is based on the fact that protein coding sequences show a specific and marked periodicity of nucleotide frequencies across the first, second and third codon positions. This periodicity serves as a robust signal for a number of gene prediction algorithms, one of the earliest and most prominent example of which is GeneMark (BORODOVSKY and MCININCH 1993). GeneMark can accurately identify protein coding nucleotide sequences based solely on the distribution of observed nucleotide frequencies across codon positions. We used the eukaryotic version of GeneMark (BORODOVSKY and MCININCH 1993), to evaluate the coding capacity of TE-derived exon sequences in the human genome. First, we compared the locations of 14,802 consensus CDS (CCDS) genes mapped to the hg17 build, from the UCSC Genome Browser (KAROLCHIK *et al.* 2003), of the human genome

to the locations of annotated TEs (see Methods). 761 of the human CCDS genes have TE-derived exon sequences; there are a total of 817 TE-derived exons. The 761 human genes with TE-derived exons include 160 TE-derived fragments with the minimum length of 100 nt required for GeneMark analysis. Using GeneMark probabilistic models (see Methods), we analyzed the TE-derived exon sequences as well as 500 randomly chosen representative non TE-derived exons by calculating their probability to be protein coding regions. The distributions of protein coding potentials (probabilities) for TE versus non TE sequences are shown in Figure 3.4. Visually the distributions are quite distinct, with TE-derived exons having far lower coding potential, and accordingly there is a highly significant difference between the two coding probability distributions, D=0.67 P=0 Kolmogorov-Smirnov test (Figure 3.4A). The average coding potential of TE-derived exons was 0.26 compared to 0.70 for non TE-derived coding sequences. Using a more sensitive custom-trained GeneMark model gave consistent results, 0.35 average TE coding probability versus 0.73 for non TE sequences with significantly different distributions D=0.59 P=0 Kolmogorov-Smirnov test (Figure 3.4B). Clearly, TE-derived exons have much lower coding probability than non TE-derived sequences suggesting that many of these exons do not actually encode proteins. Since the TE-derived exons evaluated using GeneMark as described above are taken from the RepeatMasker annotations on the human genome sequence, they do not include more ancient well established cases of TE-derived CDS such as the first three cases shown in Table 3.6. One would expect that these TE-derived CDS have higher protein coding potentials than the more recently exonized TE sequences revealed by RepeatMasker. In fact, when analyzed using GeneMark in the same way as described for the entire set of TE-derived

exons, all of their protein coding probabilities are significantly greater (z-test: $15.5<z<16.9$) than the average protein coding probability (0.35) of the aforementioned set of TE-derived exons: Telomerase=0.81, RAG1=0.77, CENP-B=0.89. Interestingly, the protein coding probability of the relatively recent case of TE-CDS exaptation, SETMAR (0.67), is also significantly greater (z=11.8) than the average coding potential for the set of RepeatMasker identified TE-derived exons. This is consistent with the fact that, while SETMAR does represent a recent case of TE-CDS exaptation, the particular TE-sequence that was exonized was already a protein-coding domain prior to becoming a host gene (CORDAUX *et al.* 2006).

Taken together, these protein coding probability data are consistent with previous studies that have suggested caution is warranted when extrapolating genome sequence analyses to infer TE-CDS exaptation events (GOTEA and MAKALOWSKI 2006; KRIEGS *et al.* 2005; PAVLICEK *et al.* 2002; WILSON *et al.* 2006). In particular, the notion that non-autonomous TEs that do not encode any protein, including SINEs such as the Alu family of elements, can emerge as protein coding sequences after being incorporated into exons has been directly challenged (PAVLICEK *et al.* 2002). On the other hand, Alus are frequently incorporated into mRNAs as exons (DAGAN *et al.* 2004; MAKALOWSKI *et al.* 1994; SOREK *et al.* 2002; YULUG *et al.* 1995), and there are a number of specific cases of Alu-derived CDS that have been proposed to provide novel CDS to primate genes (KRULL *et al.* 2005; SINGER *et al.* 2004). In light of this controversy, we have specifically evaluated the potential coding capacity of Alu-derived exons using GeneMark.

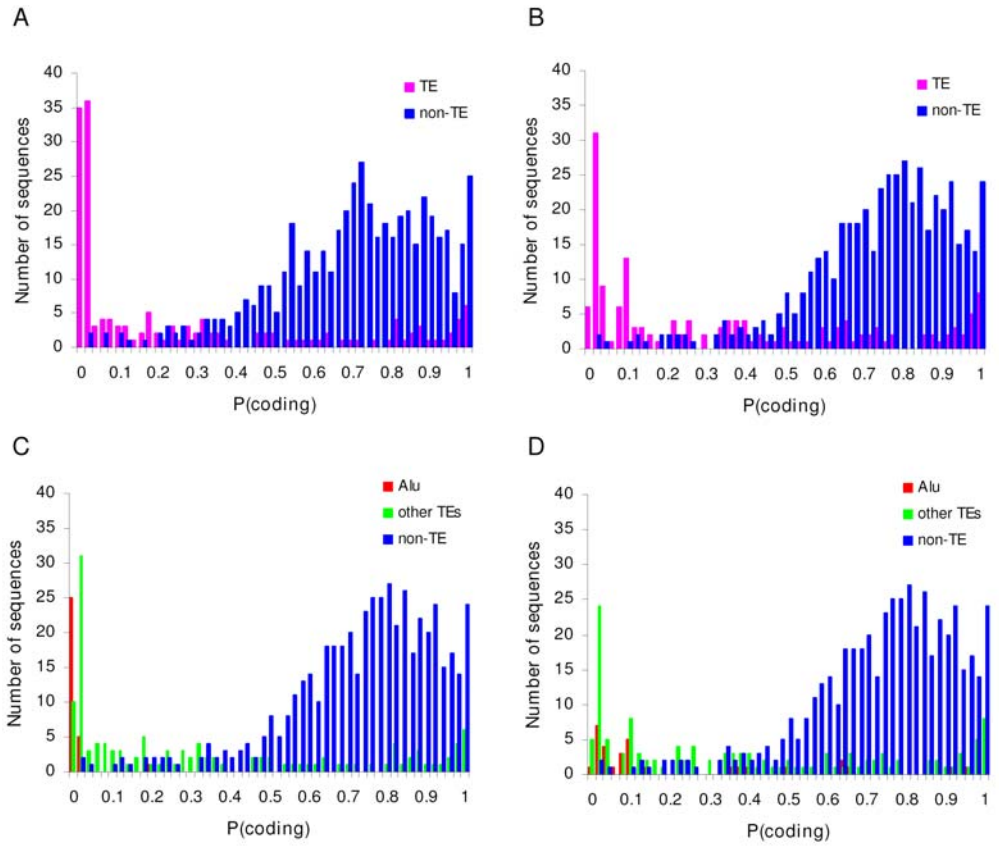

Figure 3.4: Coding probability of human CCDS genes. The coding probability of TE-derived coding sequences (pink) and non TE-derived coding sequences (blue) are shown, with results from the original GeneMark model (A) and our custom trained GeneMark model (B). TEs are separated in Alu (red) and non-Alu (green) for the original (C) and custom (D) GeneMark models.

Alu-derived exons were considered separately from all other TE-derived exons and their coding probability distributions were plotted along with the distribution for non TE-derived exons (Figure 3.4C and Figure 3.4D). Alu-derived exons have coding probability distributions that are shifted to the left, *i.e.* towards lower probability, than all other TE-derived exons. Indeed, the average coding probabilities for Alu-derived exons are significantly lower than the averages for all other TE-derived exons (Table 3.7). This result holds under a number of different analytical conditions (see Methods), including

the two different GeneMark models and the consideration of Alu-derived exons as only containing Alu sequences or containing Alu plus other TE sequences (composite TE-exons in Table 3.7).

Table 3.7: Comparison of protein coding potential for Alu-derived exons versus other TE-derived exons
Average protein coding potentials are compared between the specific pairs of groups indicated using the Student's t-test. Comparisons were done using two GeneMark models: pre-trained and custom-trained (see Methods).

GeneMark model	Comparison groups	mean	df	*t*	*P*
Pre-trained	Alu-exons *vs.* other TE-exons	0.0069 *vs.* 0.3229	158	9.8	5.2e-18
	Alu-containing composite TE-exons *vs.* other TE-exons	0.0135 *vs.* 0.3417	158	9.6	1.3e-17
Custom-trained	Alu-exons *vs.* other TE-exons	0.2034 *vs.* 0.3802	158	2.9	4.3e-3
	Alu-containing composite TE-exons *vs.* other TE-exons	0.2033 *vs.* 0.3920	158	3.4	7.4e-4

In addition to the global analysis of Alu-derived exon protein coding potential, we also evaluated several documented cases of Alu exonization events that are assumed to represent TE-CDS exaptations (KRULL *et al.* 2005; SINGER *et al.* 2004). For these cases, the specific evolutionary scenarios giving rise to the Alu-derived exons are quite well documented, but the protein coding potential of the Alu-exons appears to be assumed.

Here, the GeneMark web server (http://exon.gatech.edu/GeneMark/), which runs both GeneMark and GeneMark.hmm (LUKASHIN and BORODOVSKY 1998) programs, was used to plot protein coding probabilities along the length of the CDS using a sliding window (Figure 3.5). This allowed the protein coding potential of the Alu-derived exons

to be directly compared to that of the non TE-derived exons in the same genes. Consistent with their status as protein coding genes, the coding sequences analyzed tend to show uniformly high protein coding probabilities. However, the Alu-derived exons show far lower protein coding potential than the rest of the gene sequences. The apparent low coding potential of Alu-derived exons may also reflect the fact that these sequences have a relatively recent evolutionary origin as exons and thus have not had enough time to accumulate the kinds of changes that would yield periodicities that more closely resemble other coding sequences.

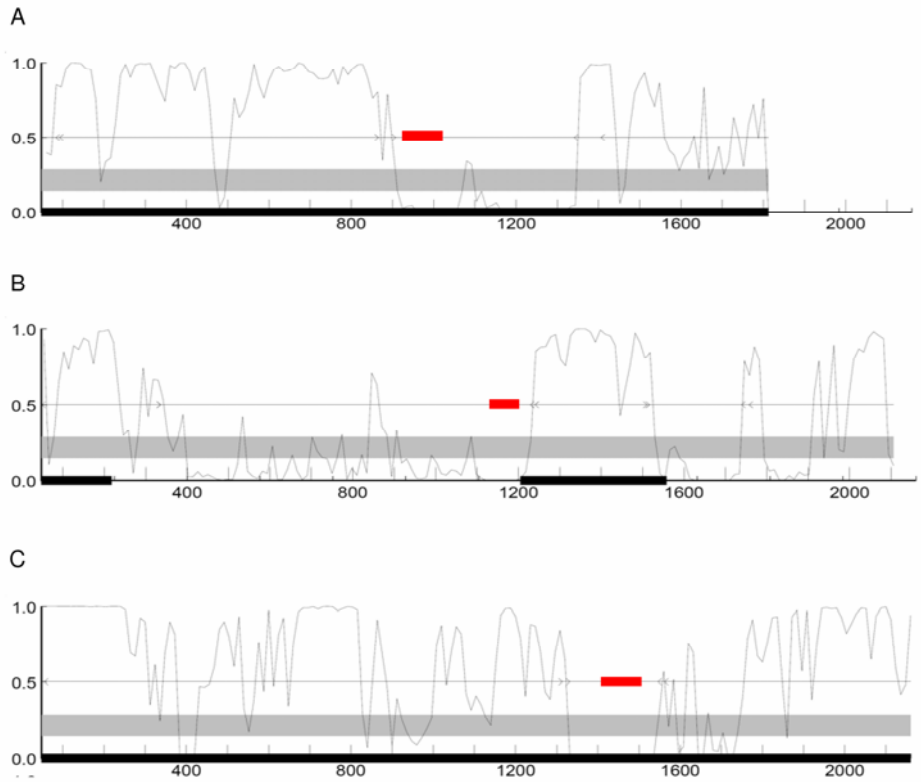

Figure 3.5: Coding probability of genes with Alu-derived exons. GeneMark protein coding probability analyses are shown for three genes with well characterized Alu-derived exons: C-rel-2 [CCDS: CCDS1864.1] (A) MTO1-3 [CCDS: CCDS4979.1] (B) and PKP2b-4 [CCDS: CCDS8731.1] (C) (KRULL *et al.* 2005). Coding probabilities were calculated within windows sliding along the length of the genes. The locations of the Alu-derived exons are shown in red.

GC codon distribution for TE-derived exons

The distribution of GC content across codon positions can also be used to evaluate the protein coding potential of genomic sequences. This kind of analysis is based on the fact that the GC level (%G+C) is distinctly lower in the second (GC2) than in the third (GC3) codon positions for protein coding sequences in species ranging from human to *Escherichia coli* (CRUVEILLER *et al.* 2007; JABBARI *et al.* 2004). Thus, for protein coding sequences, regression analysis of %GC2 x %GC3 should yield a trend line with a slope y<<1. Here, we used GC2/GC3 regression analysis to compare the protein coding potential of TE-derived versus non TE-derived exons.

For the first analysis, GC2/GC3 trends were computed for entire genes that contain one or more TE-derived exons versus entire genes with no TE-derived exons (Figure 3.6A). In this case, the GC2/GC3 distributions are indistinguishable and do not have significantly different slopes (t=0.36, df=14,798, P=0.71). However, 27.93% of TE associated genes were located outside the 95% confidence band of non TE-associated gene set. On the other hand, when TE-derived exons are considered alone (Figure 3.6B), the slopes of the TE-derived versus non TE-derived sets are significantly different (t=2.84, df=14,384, P=4.6e-3), and 31.70% of TE-derived exons are found outside the 95% confidence interval for the non TE-derived set. Thus, while the GC2/GC3 analysis appears to suffer from a lack of resolution compared to the GeneMark coding potential analysis, it too points to a relatively low coding probability for TE-derived exons.

We also analyzed Alu-derived exons separately using GC2/GC3 codon analysis as was done with GeneMark. Visual inspection of the location of Alu-derived exons on the GC2/GC3 plot shows that they have relatively higher GC2, typical of non-coding

sequence, and 41.79% fall outside the 95% confidence interval, all of which fall above the upper confidence interval boundary (See Figure B.1). In addition, Alu-derived exons have average GC2/GC3 ratios that are significantly higher than the GC2/GC3 ratios for all other TE-derived exons and for the non TE-derived gene set (Table 3.8). In other words, the GC2/GC3 analysis also suggests that Alu-derived exons are less likely to encode protein sequences than other TE-derived exons.

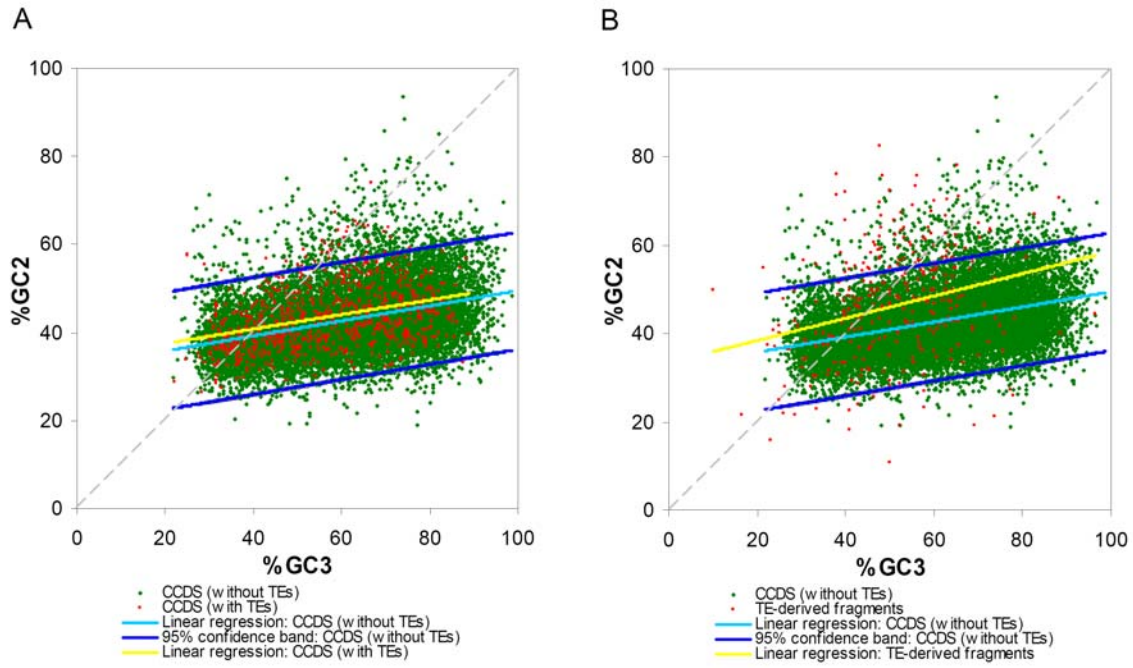

Figure 3.6: The GC composition of human CCDS genes. The scatter plots of %G+C of second (GC2) versus third (GC3) codon positions for TE-associated genes (red) and non TE-associated genes (green) are shown. The light blue line represents the linear regression line of non TE-associated genes while the blue lines show the 95% confidence interval. For the TE-associated group, the GC content for the whole sequence of TE-associated genes (A) and for TE-derived gene fragments only (B) are shown. The yellow line represents the linear regression line of these TE associated groups.

Table 3.8: Comparison of GC2/GC3 ratios for different classes of TE-derived and non TE genes (exons)
Average GC2/GC3 ratios are compared for pairs of groups indicated using the Student's t-test.

Comparison groups	averages	df	*t*	*P*
TE-genes *vs.* non TE-genes	0.82 *vs.* 0.76	14800	7.4	1.2e-13
TE-exons *vs.* non TE-genes	0.96 *vs.* 0.76	14386	9.4	6.8e-21
Alu-exons *vs.* other TE-exons	1.01 *vs.* 0.95	345	1.9	6.2e-2
Alu-exons *vs.* non TE-genes	1.01 *vs.* 0.76	14106	8.2	3.9e-16

CONCLUSIONS

The potential for TE sequences to become exapted as host protein coding sequences through the process of exonization has received a great deal of attention as of late (BOWEN and JORDAN 2007; MAKALOWSKI and TODA 2007; PIRIYAPONGSA *et al.* 2007b; VOLFF 2006; VOLFF and BROSIUS 2007). Implicit in much of this literature is the assumption that exonized TE nucleotide sequences, *i.e.* TE sequences that are spliced into mRNAs, actually encode protein sequences. However, this assumption has been challenged on several different fronts (GOTEA and MAKALOWSKI 2006; KRIEGS *et al.* 2005; PAVLICEK *et al.* 2002; WILSON *et al.* 2006). In particular, it is unclear whether non-autonomous TEs that do not encode any protein, such as Alu elements, actually provide protein coding sequences after becoming exonized (PAVLICEK *et al.* 2002). Nevertheless, recent studies continue to turn up numerous apparent cases of TE-CDS exaptation (BRITTEN 2006; ZDOBNOV *et al.* 2005). So the matter of TE-CDS exaptation remains unsettled, and in this report we have tried to address the issue from two perspectives: i-with respect to the ascertainment biases that arise from the use of different sequence similarity search methods and ii-in terms of the protein coding potential revealed by the probabilistic analysis of exonized TE nucleotide sequences.

Our use of profile-based (HMM) sequence similarity searches did allow for greater sensitivity than the more widely used DNA-DNA (*e.g.* RepeatMasker) search methods when employed on a test set of well-characterized exapted TE-CDS (Table 3.5 and Table 3.6). Thus, ascertainment biases could explain the paucity of reliable examples of TE-derived protein coding sequences uncovered via the analysis of experimentally characterized protein sequence data sets (GOTEA and MAKALOWSKI 2006; PAVLICEK *et al.* 2002). However, when profile-based search methods are similarly applied to large-scale datasets of experimentally characterized proteins, they did not turn up many more cases than previously found (Tables 3.1, Table 3.2 and Table 3.4). In fact, the profile-based search method appeared to be less sensitive than all BLAST-based search methods – nucleotide, protein or translated (Table 3.1 and Table 3.2). This apparent lack of power can actually be attributed to the superior selectivity of the profile-based methods (Figure 3.1) and suggests that many of the putative TE-CDS exaptation events turned up in BLAST searches may be spurious. In other words, profile-based search methods possess a valuable combination of sensitivity, measured by their ability to recover positive control test cases, and selectivity than any of the other search methods used. Nevertheless, the different search methods are complementary to the extent that combined search approaches are needed to thoroughly check any data set for all potential TE-CDS associations. Different search methods will also be more or less appropriate depending on the kind of exonization event that is being analyzed; for instance, it will not be possible to search for the contribution of non-coding TEs to exapted protein domains using profile methods based on protein sequence alignments.

The codon based analysis of exonized TE sequences suggests that many, if not

most, of these sequences do not actually encode any protein. Non-coding TEs that are exonized, such as Alu, have particularly low protein coding probabilities. The lack of protein coding potential does not mean that exonized TE sequences are necessarily non-functional. They may in fact play a role in post-transcriptional gene regulation. We hypothesize that many exonized TE sequences serve as natural anti-sense transcripts, which can function as double stranded RNA regulators of gene (protein) expression. The repetitive dispersed nature of exonized TE sequences may provide a mechanism by which they can serve as master regulators with influence over the expression of numerous genes throughout the genome.

METHODS

Detection of TE-encoded protein fragments

Sequence data sets

The set of functionally well characterized proteins was taken from two databases: Protein Data Bank (PDB) (BERMAN *et al.* 2000) (downloaded on 03/02/07) and Swiss-Prot Protein Database (BOECKMANN *et al.* 2003) (version 52.0). For the Swiss-Prot entries, only directly sequenced proteins were included in the data set. These directly sequenced proteins are the proteins whose amino acid sequence has been partially or completely determined experimentally by Edman degradation or by mass spectrometry and can be found by searching the Swiss-Prot database with the keyword 'Direct Protein Sequencing'. The data set of experimentally characterized protein sequences from PDB and Swiss-Prot was then filtered to remove the sequences from viruses. The nucleotide coding sequences (CDS) corresponding to the final set of protein sequences was obtained

from EMBL CDS database (http://www.ebi.ac.uk/embl/cds/). It should be noted that PDB entries can contain more than one distinct protein sequence (chain) and the same protein sequence (chain) may be found in more than one PDB entry. A data set of protein sequences encoded by TEs and their corresponding CDS sequences were extracted from Repbase (JURKA 2000) version 12.02. The data set of all TE nucleotide sequences (including non-autonomous TEs) was retrieved from Repbase website (http://www.girinst.org/repbase/).

Identification of TE-related protein domains

Protein domains that are associated with TEs were identified in version 21.0 of the Pfam database (SONNHAMMER *et al.* 1997) and the associated InterPro annotation (APWEILER *et al.* 2001). Pfam entries, both keywords and domain descriptions, were automatically searched using a set of related terms (*e.g.* transposon, retrotransposon, retroviral/ retrovirus, transposase, reverse transcriptase, etc.) as in (ZDOBNOV *et al.* 2005). The resulting putative TE-related Pfam entries were then manually inspected to remove spurious hits corresponding to protein families that are not encoded by any TEs. Manual inspection was done using the Pfam domain descriptions and literature references. HMM profiles, representing the site-specific sequence variation, of the TE-related Pfam domains were used in searches with the HMMER program as described below.

Sequence similarity searches

The experimentally characterized PDB/Swiss-Prot protein sequence data sets described above were searched for the presence of the TE-related protein domains using version 2.3.2 of the HMMER program (http://hmmer.janelia.org/). HMMER searches were run using a series of increasingly stringent threshold E-values, from E-

value≤1 to E-value≤0.00001, in addition to the gathering threshold (GA) and trusted cutoff (TC) threshold values (Table 3.1 and Table 3.2). The GA and TC threshold cutoffs are values that have been bench-marked by the developers of HMMER to ensure that a minimum number of false-positive hits are detected. The GA thresholds are empirically set for each Pfam model and correspond to the score used to collect all of the sequences included in the Pfam full alignment. In other words, the GA threshold corresponds to the complete absence of false-positives. The TC threshold is similar to GA in the sense that it corresponds to the lowest scoring hit to any sequence included as a true member of a particular Pfam domain. TE-associated PDB/ Swiss-Prot proteins detected by HMMER were classified into five categories: i-potential TE-related proteins (the host proteins containing TE-associated protein domains), ii-viral proteins (genuine viral proteins though the PDB source organism is not listed as a virus), iii-TE-encoded proteins found in TEs as opposed to cellular host proteins, iv-synthetic construct (synthesized protein sequences), and v-ubiquitous non-specific TE-related protein domains (*i.e.* host protein containing Pfam domains which are not specific to TE protein sequences but can be found in TEs as well).

Various BLAST programs (ALTSCHUL *et al.* 1990) and the program RepeatMasker (SMIT *et al.* 1996-2004) were used to search the protein sequence and CDS data sets described above for TE-related protein sequences and/or TE-related CDS. The specific program-query-database combinations used for each search are shown in Table 3.3. BLAST programs were run using a series of E-value thresholds, from E-value≤1 to E-value≤0.00001, with default parameters and without low-complexity filtering.

The fraction (*f*) hits shared between any two methods was taken as the ratio of

the number of hits retrieved in both searches to the total number of hits in both searches. For two searches that return *x* and *y* numbers of hits respectively:

$$fxy = \frac{x \bigcap y}{x + y - x \bigcap y}$$

All pairwise similarity values were calculated in this way, and the resulting matrix was clustered using hierarchical clustering. Matrix clustering and visualization were done using the programs Genesis (STURN *et al.* 2002) and Matrix2png (http://bioinformatics.ubc.ca/matrix2png/) respectively.

Analysis of known cases of TE-derived proteins (genes)

Several well known cases of proteins (genes) derived from TEs were evaluated by the HMMER, BLAST and RepeatMasker programs to determine the efficiency of different search methods in detecting TE-CDS exaptation events. The TE sequence data set sources as described in the previous section were used for these searches. The Genbank sequence accessions for the known cases are Telomerase Reverse Transcriptase [RefSeq: NM_198253, NM_198255, NP_937983, NP_937986], Recombination Activating Gene 1 (*RAG1*) [RefSeq: NM_000448, NP_000439], Centromere protein B (*CENPB*) [RefSeq: NM_001810, NP_001801], SET domain and Mariner transposase fusion gene (*SETMAR*) [RefSeq: NM_006515, NP_006506].

Evolutionary analysis of TE-associated protein domain

We used the THAP domain-containing protein, *C. elegans* C-terminal binding protein (CtBP) [PDB: 2jm3], for a phylogenetic analysis of THAP domain shared between TE and cellular proteins. The position of the THAP domain in *C. elegans* CtBP [RefSeq: NP_508983] was identified using HMMER program. The BLASTP program

was used to search for the homologous sequences of CtBP THAP in other species, using the Genbank non-redundant database, and the sequence fragments corresponding to the THAP domain were extracted as "cellular THAP". The library of TE proteins sequences (described in the sub-section Detection of TE-encoded protein fragments: Sequence data sets) was searched for the THAP-containing entries by using HMMER program with gathering (GA) threshold cutoff. The sequence fragments corresponding to the THAP domain in TE proteins were extracted as "TE-THAP" sequences.

Phylogenetic analysis of THAP sequences was done using the neighbor joining algorithm (SAITOU and NEI 1987) implemented in the MEGA program (KUMAR *et al.* 2004). Two sources of pairwise distances were used based on i-global sequence alignment of THAP domains with CLUSTALW (THOMPSON *et al.* 1994) and ii-local alignment of THAP domains using all-against-all pairwise BLASTP. For the global THAP domain sequence alignments, Poisson distances were used, and for the local THAP domain comparisons, p-values (proportion of differences) taken from the BLAST output were transformed into gamma distances using α=2.25 (NEI and KUMAR 2000). Bootstrap analysis, based on 1,000 replicates, was performed on the global THAP sequence alignment.

Codon based analysis of TE-derived exons

The UCSC Genome Brower (KAROLCHIK *et al.* 2003) and Table Browser tools (KAROLCHIK *et al.* 2004) were used to search for human protein coding sequences co-located with TEs. Genomic locations of the CCDS genes mapped to the hg17 (NCBI Build 35) version of the human genome sequence were compared to the locations of TEs annotated with the RepeatMasker program (SMIT *et al.* 1996-2004). The CCDS gene data

set (http://www.ncbi.nlm.nih.gov/CCDS/) was chosen because it represents a highly reliable set of gene models that are built from multiple lines of evidence and undergo quality analysis across several genomic centers before being released. Two data sets were created in this way: i-genes containing TE-derived exon sequences and ii-genes without TE-derived exons.

Version 2.5f of the GeneMark program (BORODOVSKY and MCININCH 1993) was used to compare the protein coding probabilities of TE-derived and non TE-derived human exons. GeneMark uses three-periodic inhomogeneous Markov models to analyze protein coding sequences and we used two models in our analysis. The first model is the GeneMark model pre-trained on validated coding and non-coding sequences of the human genome. This model is made available with the program. We also trained a customized GeneMark model using protein coding exon sequences from the non TE-derived gene set for the coding training set and intron sequences from the same genes as the non-coding training set. Each training set was classified into five groups based on %GC content (<41, 41-47, 47-53, 53-59, >=59) for separate training of the fifth order Markov chain models. Note that 100 non TE-derived genes of each GC level were randomly selected as a set of non TE test sequences and removed from the training set before model training. The GeneMark program was run on the set of genes with TE-derived exons using the custom made model parameters corresponding to the GC content of each gene. The sliding window size was chosen to be 96 nt long and the step size to be 3 nt. The average posterior probability, which characterizes the probability that the sequence encodes a protein, was calculated for each TE-derived exon sequence fragments (>100 nt) using the following formula:

$$P(cod_{x..y}) = \frac{1}{n}\sum_{i} P(cod_1|F)$$

where $x+W/2<=i<=y-W/2$, x=start position of TE fragment, y=end position of TE fragment, n=# of sliding windows for which the midpoint lies within the range of $x+W/2$ to $y-W/2$, i=the midpoint of each sliding window, $P(cod_1|F)$=posterior probability of the event that given the fragment F, it carries genetic code in frame 1 (starting from the very first nucleotide), W=the width of sliding window. The coding probability was calculated in the same way for the non TE test sequences. The analysis was repeated for the same test set using the pre-trained GeneMark models for human genome.

For the GC2/GC3 analysis, the GC level (%G+C) of second (GC2) and third (GC3) codon positions were calculated for each coding sequence of both the TE-derived and non TE-derived gene data sets. In addition, %GC2 and %GC3 were calculated for TE-derived fragments that are at least 60nt long.

ACKNOWLEDGEMENTS

The authors wish to thank Alexandre Lomsadze for guidance on the use of the GeneMark program. JP is supported by the Ministry of Science and Technology of Thailand. IKJ is supported by the School of Biology at the Georgia Institute of Technology. MB is supported by the US National Institutes of Health. The authors are most grateful to the reviewer's of the manuscript. We appreciate and value their time and expertise along with the insightful comments and questions they provided.

CHAPTER 4

ORIGIN AND EVOLUTION OF HUMAN MICRORNAS FROM TRANSPOSABLE ELEMENTS

ABSTRACT

We sought to evaluate the extent of the contribution of transposable elements (TEs) to human microRNA (miRNA) genes along with the evolutionary dynamics of TE-derived human miRNAs. We found 55 experimentally characterized human miRNA genes that are derived from TEs, and these TE-derived miRNAs have the potential to regulate thousands of human genes. Sequence comparisons revealed that TE-derived human miRNAs are less conserved, on average, than non TE-derived miRNAs. However, there are 18 TE-derived miRNAs that are relatively conserved, and 14 of these are related to the ancient L2 and MIR families. Comparison of miRNA *vs*. mRNA expression patterns for TE-derived miRNAs and their putative target genes showed numerous cases of anti-correlated expression that are consistent with regulation via mRNA degradation. In addition to the known human miRNAs that we show to be derived from TE sequences, we predict an additional 85 novel TE-derived miRNA genes. TE sequences are typically disregarded in genomic surveys for miRNA genes and target sites; this is a mistake. Our results indicate that TEs provide a natural mechanism for the origination of miRNAs that can contribute to regulatory divergence between species as well as a rich source for the discovery of as yet unknown miRNA genes.

INTRODUCTION

MicroRNAs (miRNAs) are small, ~22 nt long, noncoding RNAs that regulate gene expression (AMBROS 2004). In animals, miRNA genes are transcribed into primary miRNAs (pri-miRNAs) and processed by Drosha to yield ~70-90 nt pre-miRNA transcripts that form hairpin structures. Mature miRNAs are liberated from these longer hairpin structures by the RNaseIII enzyme Dicer (BARTEL 2004). Drosha acts in the nucleus, cleaving the pri-miRNA near the base of the hairpin stem to yield the pre-miRNA sequence. The pre-miRNA is then exported to the cytoplasm where the stem is cleaved by Dicer to produce a miRNA duplex. One strand of this duplex is rapidly degraded and only the mature ~22 nt miRNA sequence remains. The mature miRNA associates with the RNA-induced silencing complex (RISC), and together the miRNA-RISC targets mRNAs for regulation. miRNA target specificity is determined by partial complementarity with the 3' untranslated region (UTR) sequence of the mRNA, and regulation is achieved by translational repression and/or mRNA degradation. miRNAs have been implicated in a variety of functions, including developmental timing (LEE *et al.* 1993; REINHART *et al.* 2000), apoptosis (BRENNECKE *et al.* 2003), and hematopoetic differentiation (CHEN *et al.* 2004).

miRNAs were first discovered in *Caenorhabditis elegans* through genetic analysis of developmental mutants (LEE *et al.* 1993). The small RNA product of the *lin-4* gene was found to negatively regulate *lin-14* expression via interaction with a complementary region in the *lin-14* 3'-UTR. This system appeared to be unique until a second example of a similar small regulatory RNA in *C. elegans*, *let-7*, was discovered 7 years later (REINHART *et al.* 2000). Shortly thereafter, *let-7* homologs and transcripts were detected

among a phylogenetically diverse set of animals (PASQUINELLI *et al.* 2000). The realization that miRNAs represent a distinct, coherent, and abundant class of regulatory genes was finally crystallized in 2001 with the publication of three back-to-back articles in Science, reporting the discovery of numerous novel miRNA genes (LAGOS-QUINTANA *et al.* 2001; LAU *et al.* 2001; LEE and AMBROS 2001). These articles introduced the term miRNA to refer to all small RNAs with similar genomic features but unknown functions, and miRNAs have now been found in all metazoans surveyed for their presence (BARTEL 2004).

Given their relatively recent discovery and characterization, a number of open questions concerning the function and evolution of miRNAs remain. In particular, the evolutionary origins of miRNAs are not well appreciated. For instance, many miRNA genes were found to be evolutionarily conserved and this was thought to be a general characteristic of miRNAs. However, a number of non-conserved miRNAs have been recently discovered (BENTWICH *et al.* 2005). The extent to which miRNA genes evolve as paralogous gene families is also unknown. Even the upper bound on the number of miRNA genes encoded by any given genome is not known (BEREZIKOV *et al.* 2006), and the number of new entries in the miRBase registry of miRNA genes continues to grow steadily (GRIFFITHS-JONES *et al.* 2006).

We sought to evaluate the contribution of transposable elements (TEs) to the origin and evolution of human miRNA genes. Another class of regulatory RNAs, small Interfering RNAs (siRNAs), are known to be related to TEs. Interestingly, this has been pointed out as a distinction between miRNAs and siRNAs, which are closely related in

terms of structure, function, and biogenesis. As opposed to siRNAs, miRNAs were thought to derive from loci distinct from other genes or TEs (BARTEL 2004). However, several examples of miRNA genes that are derived from TEs have been recently identified (BORCHERT *et al.* 2006; PIRIYAPONGSA and JORDAN 2007; SMALHEISER and TORVIK 2005). We wanted to look at this phenomenon more closely to identify the full extent of human miRNA genes that are related to TEs and to characterize how these genes evolve as well as their regulatory and functional potential.

TEs have several characteristics that make them interesting candidates for donating miRNA sequences. First of all, TEs are ubiquitous and abundant genomic sequences. Thus, they could provide for the emergence of paralogous miRNA gene families as well as multiple target sites dispersed throughout the genome. Since TEs tend to be among the most rapidly evolving of all genomic sequences, they may also provide a mechanism for the emergence of lineage-specific miRNA genes that could exert diversifying regulatory effects. Finally, the full contribution of TEs to miRNA sequences is likely to be underestimated due to ascertainment biases. This is because computational methods aimed at the detection of novel miRNAs tend to purposefully exclude TE sequences (BENTWICH *et al.* 2005; LI *et al.* 2006; LINDOW and KROGH 2005; NAM *et al.* 2005). This is often done for reasons of tractability, but also reflects the widely held notion that TEs are genomic parasites that do not play any functional role for their host species (DOOLITTLE and SAPIENZA 1980; ORGEL and CRICK 1980). However, many studies have identified a variety ways in which TEs have been domesticated (MILLER *et al.* 1992) to provide functions to their hosts (KIDWELL and LISCH 2001). These cases include the donation of coding sequences (VOLFF 2006) as well as numerous instances of

TE-derived regulatory sequences (BRITTEN 1996; JORDAN *et al.* 2003; VAN DE LAGEMAAT *et al.* 2003).

To evaluate the contribution of TEs to human miRNAs, we compared the genomic locations of TEs to the locations of experimentally validated human miRNA sequences reported in the miRBase database (GRIFFITHS-JONES *et al.* 2006). The evolutionary dynamics of TE-related miRNAs were evaluated by within- and between-genome sequence comparisons. The potential regulatory and functional significance of TE-derived miRNAs was explored by combining information on miRNA target site prediction, expression data for miRNA-mRNApairs, and gene functional annotations. We also sought to discover putative cases of novel TE-derived miRNA genes in the human genome through *ab initio* prediction.

MATERIALS AND METHODS

Detection

Human miRNA sequences and predicted target sites were taken from version 8.2 of the miRBase database (GRIFFITHS-JONES *et al.* 2006). These data do not include *ab initio* miRNA gene predictions. The UCSC Genome Browser (KENT *et al.* 2002) and Table Browser (KAROLCHIK *et al.* 2004) tools were used to search for miRNA genes co-located with TEs and to compare the evolutionary rates of miRNA genes. Human miRNA sequences were mapped to the hg18 (NCBI build 36.1) version of the human genome sequence and a generic feature format ‘‘custom track’’ was created (available upon request). Genomic locations of the miRNAs were compared to the locations of TEs annotated with the RepeatMasker program (SMIT *et al.* 1996-2004). For this purpose, pre-computed RepeatMasker annotations of hg18 were combined with RepeatMasker

determined genomic locations of a set of 96 ‘‘conserved’’ TE families recently added to Repbase (JURKA *et al.* 2005). These conserved consensus sequences correspond to low copy number TEs that show anomalously low levels of between genome orthologous sequence divergence and can be found by searching Repbase (http://www.girinst.org/) with the keyword ‘‘conserved’’. Sequences of TE-derived miRNAs were compared to the human genome sequence using BLAT (KENT 2002) criteria used for genome sequence hits were (1) ≥80% sequence identity with the query miRNA sequence and (2) the genomic hit region must be ≥80% and ≤120% of the length of the miRNA query sequence. The latter requirement was used to ensure that long genomic insertions were not identified as putative paralogous miRNAs.

Evolution

Comparative genomic sequence data from the UCSC Genome Browser were used to analyze the relative evolutionary rates of human miRNAs. Evolutionary rates were derived from multiple whole genome sequence alignments between the human and 16 other vertebrate genomes (BLANCHETTE *et al.* 2004; KENT *et al.* 2003). Human miRNA evolutionary rates were calculated in two ways: (1) by evaluating the number of conserved sites per miRNA and (2) by evaluating the per-site conservation scores of miRNA sequences. Conserved human genome sites were predicted by the phastCons program, which uses a phylogenetic hidden Markov model to calculate the probabilities of sites being either conserved or non-conserved (SIEPEL *et al.* 2005). Conservation scores for human genome sites were also taken from the phastCons analysis of the vertebrate multiple genome sequence alignment, and these scores correspond to the posterior probability that a site is conserved or non-conserved.

Regulation and function

Human miRNA target site predictions were taken from miRBase, which uses a modified protocol based on the miRanda algorithm (ENRIGHT *et al.* 2003). The locations of target site sequences in the human genome were compared to the RepeatMasker-based TE annotations. Expression levels for human miRNAs across five tissues (thymus, brain, liver, placenta, and testis) were taken from an oligonucleotide-based microarray study (BARAD *et al.* 2004). Human mRNA expression levels from corresponding mRNA targets were taken from the Novartis Symatlas data set (SU *et al.* 2004). Corresponding miRNA and mRNA expression profiles were normalized using standard *z*-score transformation with the program Spotfire (http://www.spotfire.com) and compared using the Pearson correlation coefficient. Gene expression data were visualized using the Genesis program (STURN *et al.* 2002). Gene ontology (GO) analysis (ASHBURNER *et al.* 2000) was done using the GOTree Machine program (ZHANG *et al.* 2004). GOTree Machine was used to identify significantly over-represented biological process GO terms from a set of genes predicted to be regulated by a particular miRNA and to plot the location of these GO terms along the GO directed acyclic graph.

TE-miRNA prediction

TE locations in the human genome were considered together with the output of the program EvoFold, which combines RNA secondary structure prediction with the evaluation of multiple sequence alignments to identify conserved secondary structures (PEDERSEN *et al.* 2006). TE sequences that encode conserved hairpin structures with length ≥55 bp, a single terminal loop ≤20 bp, and at least six paired bases in the stem region (BENTWICH *et al.* 2005) were chosen for further analysis. For conserved TE-

encoded hairpins of <55 bp that met all other criteria, the predicted secondary structure sequences were extended manually and rechecked for the ability to form hairpin structures using the program RNAfold from the Vienna RNA package (HOFACKER *et al.* 1994). Sequences that were able to encode hairpins ≥55 bp after manual extension were chosen for further analysis. The potential for putative TE-derived miRNAs identified in this way to be expressed was evaluated using EST and mRNA data. Our TE-miRNA prediction protocol is represented in Figure C.1.

RESULTS

Transposable element-derived miRNAs

miRBase is an online database of miRNA gene sequences and predicted target sites (GRIFFITHS-JONES *et al.* 2006); version 8.2 of miRBase contained 462 human miRNA gene sequences. Of these human miRNA genes, 379 are defined on the basis of experimental information, cloning of mature miRNA sequences for the most part, while 83 are predictions on the basis of sequence similarity with miRNAs that have been experimentally characterized in related species. We mapped these human miRNA genes to the complete genome sequence and compared their locations to the locations of annotated TEs. A total of 68 human miRNA genes share sequences with TEs, and all but 7 of these correspond to miRNAs experimentally characterized from human samples. The absence of *ab initio* miRNA gene predictions in the miRBase data set ensures that we are uncovering *bona fide* TE-miRNA relationships. Of these TE-related miRNAs, 49 are found in intron sequences while 19 are intergenic.

TE-related miRNAs differ in terms of the extent of overlap with TE sequences and the number of distinct TE sequences from which they are derived. For each

individual TE-related human miRNA, a schematic in Figure C.2 illustrates the identity of all co-located TE sequences along with the extent and position of the TE-miRNA overlap and the relationship between the strand-specific orientation of the TE and the miRNA. The majority (50 of 68) of TE-related miRNAs consist of >50% TE-derived positions (Figure 4.1A), and this figure is likely to be an underestimate since many TE sequences are known to have diverged beyond the ability to be recognized by the RepeatMasker annotation software. The TE-miRNA overlap distribution for the region of the miRNA gene that corresponds to the processed (mature) regulatory sequence is even more bimodal (Figure 4.1B); 47 sequences have >95% of mature miRNA positions covered by TE sequence. Nevertheless, there are a handful (7 of 68) of TE-related miRNA genes that have <20% of their sequences co-located with TE sequence. These may represent spurious cases of TE-miRNA overlap. Visual inspection of the TE-miRNA alignments (Figure C.2) was used to eliminate these unreliable cases. Only the 55 cases with at least 50% TE coverage of the pre-miRNA sequence and/or 100% TE coverage of the mature miRNA sequence were considered as actual TE-derived miRNAs and used for further analysis (Table 4.1). These 55 TE-derived miRNAs represent ~12% (55/462) of all human miRNAs reported in miRBase version 8.2.

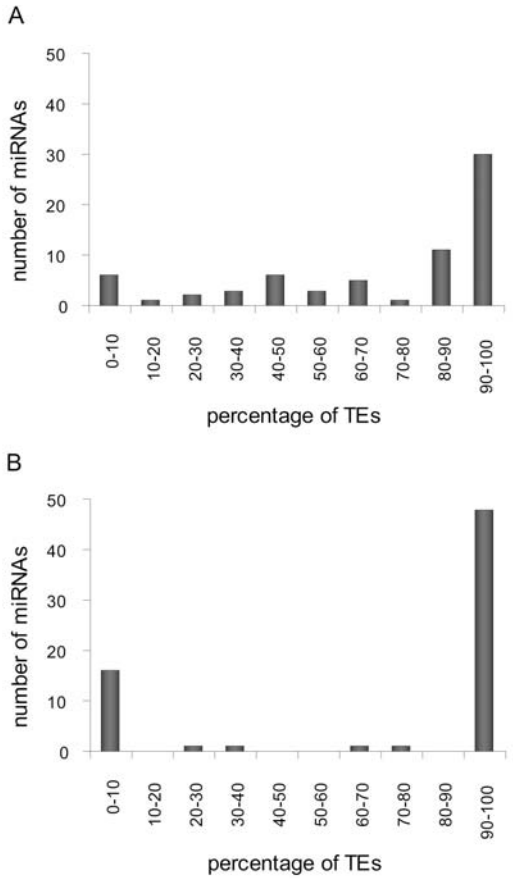

Figure 4.1: Percentage of TE-derived residues in miRNA genes. Frequency distributions are shown for the percentages of TE-derived residues relative to miRNA gene sequences (A) and mature miRNA sequences (B).

Table 4.1: TE-derived human miRNAs

Name[a]	Accn[b]	Coords[c]	TE[d]	Overlap[e]	Cons[f]	Targets[g]
hsa-mir-130b	MI0000748	chr22:20337593-20337674(+)	MIRm	65.85	0.8492	865 (10.75%)
hsa-mir-151	MI0000809	chr8:141811845-141811934(-)	L2	100.00	0.9317	863 (12.28%)
hsa-mir-28	MI0000086	chr3:189889263-189889348(+)	L2	93.02	0.9979	1136 (10.21%)
hsa-mir-325	MI0000824	chrX:76142220-76142317(-)	L2	89.80	0.9905	751 (13.32%)
hsa-mir-330	MI0000803	chr19:50834092-50834185(-)	MIRm	53.19	0.9867	927 (5.18%)
hsa-mir-345	MI0000825	chr14:99843949-99844046(+)	MIR	39.80	0.8265	895 (7.82%)
hsa-mir-361	MI0000760	chrX:85045297-85045368(-)	MER5A	81.94	0.9998	882 (14.51%)
hsa-mir-370	MI0000778	chr14:100447229-100447303(+)	MIRm	100.00	0.9893	1006 (4.77%)
hsa-mir-374	MI0000782	chrX:73423846-73423917(-)	L2	54.17	0.9970	773 (7.50%)
hsa-mir-378	MI0000786	chr5:149092581-149092646(+)	MIRb	90.91	1.0000	0 (0%)
hsa-mir-421	MI0003685	chrX:73354937-73355021(-)	L2	89.41	0.9999	1023 (14.47%)
hsa-mir-422a	MI0001444	chr15:61950182-61950271(-)	MIR3	100.00	0.0018	940 (7.34%)
hsa-mir-493	MI0003132	chr14:100405150-100405238(+)	L2	66.29	0.9990	0 (0%)
hsa-mir-513-1	MI0003191	chrX:146102673-146102801(-)	MER91C	100.00	0.0543	1065 (7.14%)
hsa-mir-513-2	MI0003192	chrX:146115036-146115162(-)	MER91C	100.00	0.0003	1065 (7.14%)

Table 4.1 continued

Name[a]	Accn[b]	Coords[c]	TE[d]	Overlap[e]	Cons[f]	Targets[g]
hsa-mir-544	MI0003515	chr14:100584748-100584838(+)	MER5A1	100.00	0.9337	1056 (10.42%)
hsa-mir-545	MI0003516	chrX:73423664-73423769(-)	L2	82.08	0.9958	1065 (16.35%)
hsa-mir-548a-1	MI0003593	chr6:18679994-18680090(+)	MADE1	78.35	0.0391	1255 (7.09%)
hsa-mir-548a-2	MI0003598	chr6:135601991-135602087(+)	LTR16A1, MADE1	100.00	0.0047	1255 (7.09%)
hsa-mir-548a-3	MI0003612	chr8:105565773-105565869(-)	MLT1G1, MADE1	100.00	0.0044	1255 (7.09%)
hsa-mir-548b	MI0003596	chr6:119431911-119432007(-)	MADE1	83.51	0.0175	1197 (5.93%)
hsa-mir-548c	MI0003630	chr12:63302556-63302652(+)	MADE1	83.51	0.0092	1302 (6.76%)
hsa-mir-548d-1	MI0003668	chr8:124429455-124429551(-)	MADE1	83.51	0.0076	1055 (10.24%)
hsa-mir-548d-2	MI0003671	chr17:62898067-62898163(-)	MADE1	83.51	0.0000	1055 (10.24%)
hsa-mir-552	MI0003557	chr1:34907787-34907882(-)	L1MD2	100.00	0.0000	1067 (11.62%)
hsa-mir-558	MI0003564	chr2:32610724-32610817(+)	MLT1C	45.74	0.0112	778 (7.58%)
hsa-mir-562	MI0003568	chr2:232745607-232745701(+)	L1MB7	100.00	0.0019	954 (11.64%)
hsa-mir-566	MI0003572	chr3:50185763-50185856(+)	AluSg	100.00	0.0000	1184 (80.07%)
hsa-mir-570	MI0003577	chr3:196911452-196911548(+)	MADE1	82.47	0.0000	1115 (4.22%)
hsa-mir-571	MI0003578	chr4:333946-334041(+)	L1MA9	96.88	0.0000	948 (8.33%)
hsa-mir-575	MI0003582	chr4:83893514-83893607(-)	MIR	61.70	0.0001	1048 (7.35%)
hsa-mir-576	MI0003583	chr4:110629303-110629400(+)	L1MB7	100.00	0.0121	921 (10.53%)
hsa-mir-578	MI0003585	chr4:166526844-166526939(+)	L2	44.79	0.0064	1012 (7.61%)
hsa-mir-579	MI0003586	chr5:32430241-32430338(-)	MADE1, L1MB8	100.00	0.3543	1202 (6.32%)
hsa-mir-582	MI0003589	chr5:59035189-59035286(-)	L3, L3	85.71	0.9954	1017 (8.06%)
hsa-mir-584	MI0003591	chr5:148422069-148422165(-)	MER81	92.78	0.0008	794 (10.96%)
hsa-mir-587	MI0003595	chr6:107338693-107338788(+)	MER115	100.00	0.0053	970 (6.39%)
hsa-mir-588	MI0003597	chr6:126847470-126847552(+)	L1MA3	100.00	0.0000	873 (10.77%)
hsa-mir-603	MI0003616	chr10:24604620-24604716(+)	MADE1	84.54	0.0102	1008 (7.44%)
hsa-mir-606	MI0003619	chr10:76982222-76982317(+)	L1MCc	100.00	0.0014	776 (8.38%)
hsa-mir-607	MI0003620	chr10:98578416-98578511(-)	MIR	100.00	0.9990	985 (8.83%)
hsa-mir-616	MI0003629	chr12:56199213-56199309(-)	L2	100.00	0.0004	922 (10.30%)
hsa-mir-619	MI0003633	chr12:107754813-107754911(-)	L1MC4, AluSx	100.00	0.0008	765 (8.89%)
hsa-mir-625	MI0003639	chr14:65007573-65007657(+)	L1MCa	100.00	0.0018	1065 (4.41%)
hsa-mir-626	MI0003640	chr15:39771075-39771168(+)	L1MB8, L1MCa	56.38	0.0086	1022 (6.65%)
hsa-mir-633	MI0003648	chr17:58375308-58375405(+)	MIRb	100.00	0.0136	843 (7.12%)
hsa-mir-634	MI0003649	chr17:62213652-62213748(+)	L1ME3A	48.45	0.0019	886 (5.08%)
hsa-mir-640	MI0003655	chr19:19406872-19406967(+)	MIRb	100.00	0.0074	853 (28.49%)
hsa-mir-644	MI0003659	chr20:32517791-32517884(+)	L1MB3	61.70	0.1035	970 (4.95%)
hsa-mir-645	MI0003660	chr20:48635730-48635823(+)	MER1B	62.77	0.0002	682 (13.49%)
hsa-mir-648	MI0003663	chr22:16843634-16843727(-)	L2	98.94	0.0008	943 (6.15%)
hsa-mir-649	MI0003664	chr22:19718465-19718561(-)	L1M4, MER8, AluSx	100.00	0.0005	1033 (10.65%)
hsa-mir-652	MI0003667	chrX:109185213-109185310(+)	MER91C	100.00	0.9883	803 (39.36%)
hsa-mir-659	MI0003683	chr22:36573631-36573727(-)	Arthur1	46.39	0.0027	890 (8.20%)
hsa-mir-95	MI0000097	chr4:8057928-8058008(-)	L2	95.06	0.9862	847 (16.06%)

[a]miRNA name (from miRBase)
[b]miRBase accession number
[c]Human genome (hg18) coordinates of the miRNA
[d]Name of co-located TE
[e]Percent of miRNA overlapping with TE sequence
[f]Average conservation score
[g]Total number of targets with percent derived from TEs shown in parentheses

The TE-related miRNAs that we identified are derived from all four major classes of human TEs: long- and short- interspersed nuclear elements (LINE and SINE), long terminal repeat containing elements (LTR) and DNA-type transposons (Table 4.1). Specific classes and families of TEs show marked over- or under-representation among human miRNAs (Figure 4.2). The related L2 (LINE) and MIR (SINE) families, as well as DNA elements, show far more overlap with miRNA genes than is expected on the basis of their relative frequency in the genome (37 observed versus 11 expected; χ^2=30.74 *P*=3.0e-8). Most of the DNA-type elements that contribute to miRNA genes are short non-autonomous derivatives of full-length transposons known as miniature inverted-repeat transposable elements (MITEs). This includes a group of seven closely related miRNA genes (hsa-mir-548), which are all derived from the Made1 family of MITEs (PIRIYAPONGSA and JORDAN 2007). Alu (SINE) elements and LTR type TEs are generally under-represented among TE-derived miRNA genes. Most TE-related miRNA genes are derived from a single TE insertion, but there are several examples where nested insertion events have led to the origin of a single miRNA gene from two or even three TEs (Figure C.2). For instance, there are two cases where a Made1 element inserted into an LTR element yielded a miRNA gene (examples 24 and 27 in Figure C.2), and an insertion of an Alu into a L1 (LINE) sequence also gave rise to a miRNA gene (example 46 in Figure C.2).

TE-derived human miRNA genes were used as queries in BLAT searches against the human genome sequence to search for putative paralogs. There are 19 cases of TE-derived miRNA genes with closely related paralogs in the human genome (Table 4.2).

The number of paralogs per miRNA ranges from 1, for the L1-derived hsa-mir-552, to 145, for the Made1-derived hsa-mir-548d-2.

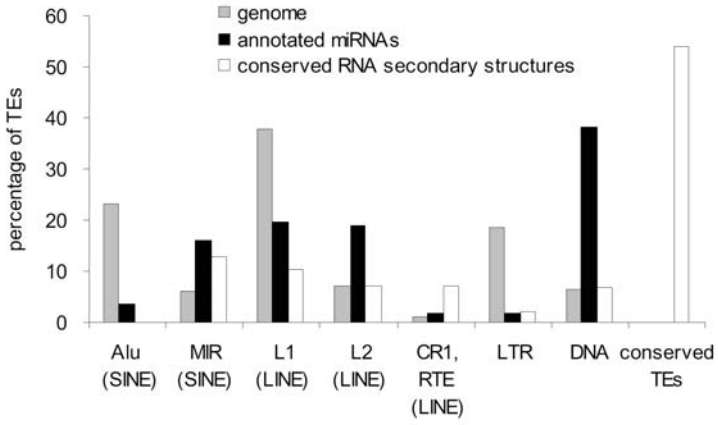

Figure 4.2: Percentage of TE sequences among different classes and families for the human genome (shading) and for TE-derived miRNA genes (solid). Relative percentages are shown such that the total will sum to 100% for the genome and for miRNAs.

Table 4.2: Putative TE-derived miRNA paralogs

Name[a]	Accn[b]	TE[c]	Paralogs[d]
hsa-mir-513-1	MI0003191	MER91C	3
hsa-mir-513-2	MI0003192	MER91C	3
hsa-mir-548a-1	MI0003593	MADE1	24
hsa-mir-548a-2	MI0003598	LTR16A1, MADE1	81
hsa-mir-548a-3	MI0003612	MLT1G1, MADE1	82
hsa-mir-548b	MI0003596	MADE1	23
hsa-mir-548c	MI0003630	MADE1	124
hsa-mir-548d-1	MI0003668	MADE1	71
hsa-mir-548d-2	MI0003671	MADE1	145
hsa-mir-552	MI0003557	L1MD2	1
hsa-mir-562	MI0003568	L1MB7	2
hsa-mir-566	MI0003572	AluSg	87
hsa-mir-570	MI0003577	MADE1	48
hsa-mir-571	MI0003578	L1MA9	4
hsa-mir-579	MI0003586	MADE1, L1MB8	3
hsa-mir-603	MI0003616	MADE1	30

Table 4.2 continued

Name[a]	Accn[b]	TE[c]	Paralogs[d]
hsa-mir-607	MI0003620	MIR	1
hsa-mir-649	MI0003664	L1M4, MER8, AluSx	4
hsa-mir-652	MI0003667	MER91C	4

[a]miRNA name (from miRBase)
[b]miRBase accession number
[c]Name of co-located TE
[d]Number of paralogous sequences in the human genome

Evolution of TE-derived miRNAs

Comparative genomic sequence data were used to assess the relative evolutionary rates of TE-derived miRNAs. This analysis was based on whole genome sequence alignments between humans and 16 other vertebrate species. Two related approaches were used to evaluate the conservation of individual miRNA sequence sites across vertebrate genomes; the first approach results in a binary characterization of either conserved or non-conserved for each site, while the second rests on a more continuous score that relates the probability of a site being conserved. All genome sites for human miRNAs were considered using these two metrics, and the relative conservation levels for TE-derived *vs.* non TE-derived miRNA genes were compared. A total of 32.1% of sites in TE-derived miRNAs map to the most conserved elements in the human genome. This is far greater than the ~5% of conserved sites seen for the entire human genome but significantly less than seen for non TE-derived miRNAs, which have 63.2% conserved sites (t=4.39 P=1.4e-5 Student's t-test) (Figure 4.3A). When the per-site conservation probabilities of human miRNAs were measured, a similar pattern was observed. The average conservation score of TE-derived miRNAs was 0.33 compared to 0.63 for non TE-derived miRNAs (t=4.37 P=1.5e-5 Student's t-test) (Figure 4.3B). In addition, the

frequency distribution of the average conservation scores for all human miRNA genes reveals that, compared to non TE-derived miRNAs, there are far more TE-derived miRNAs that show little or no conservation and fewer that are highly conserved (Figure 4.3C). Thus, on the whole, TE-derived miRNAs are significantly less conserved than non TE-derived miRNAs.

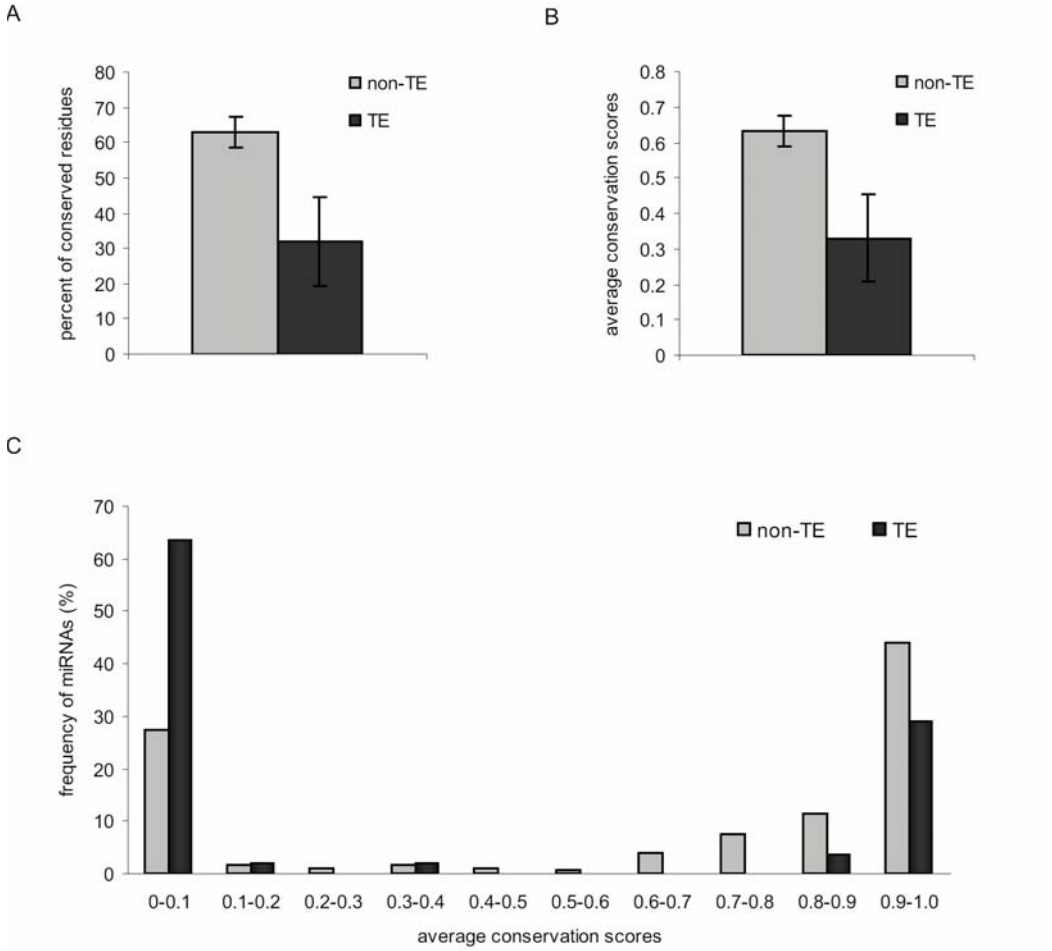

Figure 4.3: Evolutionary conservation of human miRNA genes. (A) The percentage of conserved residues for non TE-derived miRNAs (shading) versus TE-derived miRNAs (solid) with 95% confidence intervals shown. (B) The average per-site conservation score for non TE-derived miRNAs (shading) versus TE-derived miRNAs (solid) with 95% confidence intervals shown. (C) Frequency distribution of the average per-site conservation scores for non TE-derived miRNAs (shading) versus TE-derived miRNAs (solid).

We used the frequency distribution of average conservation scores to divide TE-derived miRNAs into conserved (≥0.8 average conservation probability) and non-conserved (<0.8 average conservation probability) groups. Using this criteria, there are 37 non-conserved and 18 conserved TE-derived miRNAs (Table 4.1). The least conserved TE-derived miRNAs are primate specific, having orthologous sequences in the chimpanzee only or both the chimpanzee and rhesus genomes. Of 18 conserved miRNAs, 14 are derived from the L2 and MIR families; this is far more than would be expected on the basis of the overall frequency of L2 and MIR sequences among TE-derived miRNAs (χ^2=17.8 *P*=3.6e-5). The conservation of L2 and MIR TE-derived miRNAs is consistent with a previous study that found many anomalously conserved L2 and MIR sequences (SILVA *et al.* 2003). Indeed, L2 and MIR are relatively ancient TE families with many sequences that inserted prior to the divergence of the human and mouse evolutionary lineages. We observed 10 of the conserved L2- and MIR-derived miRNA sequences to have orthologous sequences in the mouse genome, and there are 9 orthologous mouse miRNAs in these regions that are annotated in miRBase (Table 4.3). All of the 8 conserved L2 miRNAs are derived from the same region near the 3'end of the L2 consensus sequence (approximately positions 3200–3400), while the 6 MIR-derived miRNAs are found in dispersed locations on the MIR consensus sequence.

A frequency distribution of conserved *vs.* non-conserved TE-derived miRNA genes, compared to genome wide relative TE frequencies, reveals distinct conservation levels for miRNAs derived from particular TE classes/families (Figure 4.4). For instance, L2 and MIRs contribute far more conserved than non-conserved miRNAs, and the fraction of conserved L2 and MIR elements in miRNAs is much higher than seen for

these same elements in the genome as a whole. DNA-type elements show the opposite pattern. There is a higher fraction of non-conserved DNA-type elements among miRNAs than is seen for the whole genome. All of the miRNAs derived from Alu and L1 elements are non-conserved.

Table 4.3: Human-mouse orthologous miRNAs derived from L2 and MIR TEs

Human miRNA[a]	Human coords[b]	TE[c]	Mouse miRNA[a]	Mouse coords[b]
hsa-mir-345: MI0000825	chr14:99843949-99844046(+)	MIR	mmu-mir-345: MI0000632	chr12:109,284,780-109,284,874 (+)
hsa-mir-130b: MI0000748	chr22:20337593-20337674(+)	MIRm	mmu-mir-130b :MI0000408	chr16:17,037,626-17,037,705(-)
hsa-mir-151: MI0000809	chr8:141811845-141811934(-)	L2	mmu-mir-151: MI0000173	gap
hsa-mir-95: MI0000097	chr4:8057928-8058008(-)	L2	-	gap
hsa-mir-330: MI0000803	chr19:50834092-50834185(-)	MIRm	mmu-mir-330: MI0000607	chr7:18,339,991-18,340,084(+)
hsa-mir-370: MI0000778	chr14:100447229-100447303(+)	MIRm	mmu-mir-370: MI0001165	chr12:110,066,065-110,066,139(+)
hsa-mir-325: MI0000824	chrX:76142220-76142317(-)	L2	mmu-mir-325: MI0000597	chrX:101,581,801-101,581,898(-)
hsa-mir-545: MI0003516	chrX:73423664-73423769(-)	L2	-	chrX:99,818,159-99,818,260(-)
hsa-mir-374: MI0000782	chrX:73423846-73423917(-)	L2	mmu-mir-374: MI0004125	chrX:99,818,306-99,818,361(-)
hsa-mir-28: MI0000086	chr3:189889263-189889348(+)	L2	mmu-mir-28: MI0000690	chr16:24,743,204-24,743,289(+)
hsa-mir-493: MI0003132	chr14:100405150-100405238(+)	L2	-	chr12:110,028,035-110,028,123(+)
hsa-mir-607: MI0003620	chr10:98578416-98578511(-)	MIR	-	gap
hsa-mir-421: MI0003685	chrX:73354937-73355021(-)	L2	-	chrX:99,775,634-99,775,718(-)
hsa-mir-378: MI0000786	chr5:149092581-149092646(+)	MIRb	mmu-mir-378: MI0000795	gap

[a]miRBase names and accessions for human and mouse orthologous miRNAs
[b]Genome coordinates for human and mouse orthologous regions; 'gap' means no orthologous region.
[c]Name of the related TE sequence

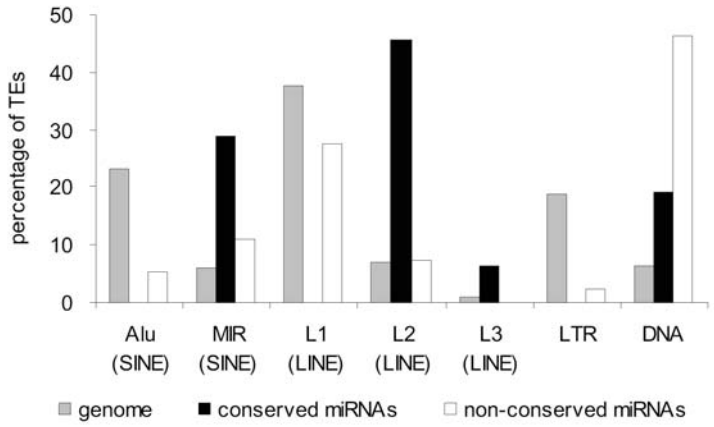

Figure 4.4: Percentage of TE sequences among different classes and families for the human genome (shading), for conserved TE-derived miRNAs (solid) and for non-conserved TE-derived miRNAs (open). Relative percentages are shown such that the total will sum to 100% for the genome and for each group of miRNAs.

Regulation and function

Given their high copy numbers, there is a potential for TE-derived miRNAs to regulate multiple genes via homologous target sites dispersed throughout genome. Using the miRBase target predictions, TE-derived miRNAs were found to have hundreds of putative target sites (Table 4.1; Figure 4.5A). However, while many of these target sites are also derived from TEs, in most cases the proportion of TE-derived target sites is ~10% (Table 4.1; Figure 4.5B). Thus, TE-derived miRNAs also have the potential to regulate host genes with non TE-derived targets. The relative paucity of TE-derived target sites can be attributed, in part, to the fact that target site prediction methods employ conservation of 3' UTR sequences as one criteria and TEs tend to be lineage specific and non-conserved.

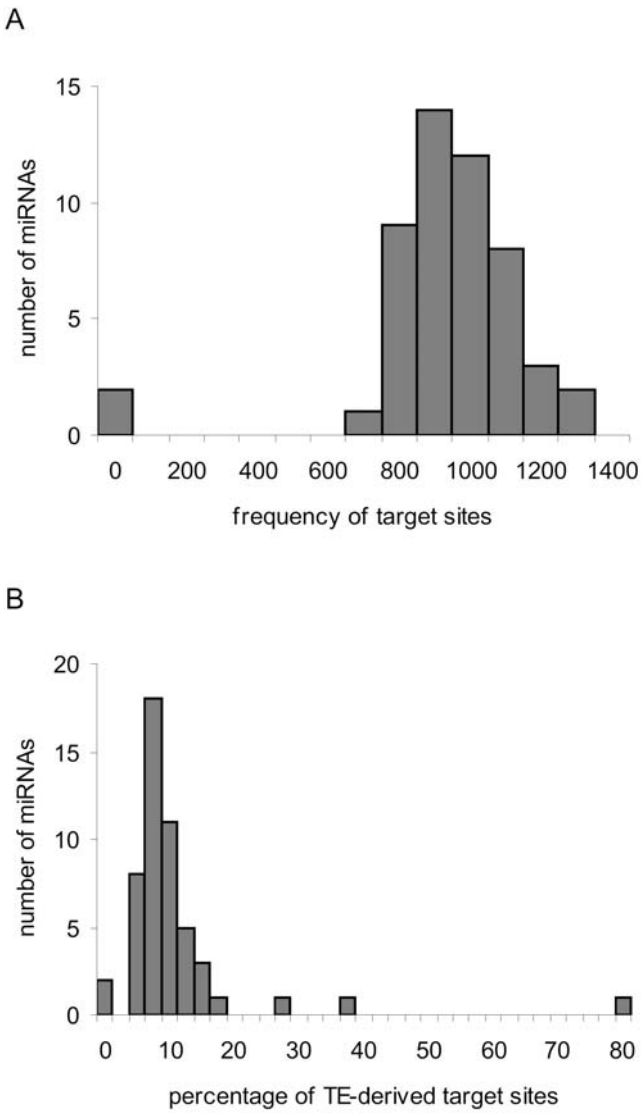

Figure 4.5: Target site frequencies for TE-derived miRNAs. (A) Frequency distribution showing the number of target sites per TE-derived miRNA. (B) Frequency distribution showing the percentage of TE-derived target sites per TE-derived miRNA.

There are several outliers that have a substantially higher fraction of TE-derived target sites. For instance, hsa-mir-566 is derived from Alu and it has 1,184 predicted targets with 948 (80%) derived from TEs. Most of these TE-derived hsa-mir-566 target sites are related to Alu insertions and this is consistent with previous studies that have

found numerous putative Alu-related miRNA target sites in the human genome (DASKALOVA *et al.* 2006; SMALHEISER and TORVIK 2006).

The predicted target sites analyzed here are all putative sites and it is difficult to know with certainty whether they are actually involved in miRNA-mediated gene regulation. Another way to evaluate the regulatory potential of miRNAs is to compare the expression patterns of miRNAs to the expression patterns of the genes they are thought to regulate (FARH *et al.* 2005; HUANG *et al.* 2006; SOOD *et al.* 2006; STARK *et al.* 2005). The rationale behind the miRNA-mRNA expression pattern comparison is based on the mRNA degradation model of miRNA action. According to this model, miRNA binding to mRNA target sites causes the mRNA transcripts to be degraded. This model predicts anti-correlations between expression levels of miRNAs and the mRNAs of their target genes; *i.e.*, high levels of miRNA would lead to decreased levels of targeted mRNA.

We sought to compare miRNA expression levels for TE-derived miRNA genes to mRNA expression levels of their target genes to look for anti-correlations that are consistent with regulation via mRNA degradation. miRNA expression data were taken from a microarray study of 150 human miRNAs across five tissue samples (BARAD *et al.* 2004), and mRNA expression data were taken from the Novartis SymAtlas (SU *et al.* 2004). Pairs of miRNA-mRNA gene expression profile vectors were compared using the Pearson correlation coefficient (r). There were only three TE-derived miRNA genes with expression data available. Despite this small sample size and the fairly low resolution afforded by the comparison of only five tissues, we found numerous cases of strongly anti-correlated miRNA-mRNA pairs (Figure 4.6). Since this anti-correlation is consistent with the mRNA degradation model of miRNA gene regulation, it provides an additional

source of support for putative miRNA target sites and the regulatory action of TE-derived miRNAs.

We also evaluated the GO biological process annotations of the anti-correlated gene sets to look for over-represented functional categories that may indicate specific functional roles for TE-derived miRNAs. The top 10% of anti-correlated mRNAs (*i.e.*, those with the lowest *r*-values) for each of the three TE-derived miRNAs with expression data were evaluated for over-represented GO terms. The miRNA hsa-mir-130b gave the strongest signal of GO term over-representation; 39 of 80 genes were found to correspond to significantly over-represented GO terms (Table C.1). Many of these genes correspond to metabolism and transcriptional regulation in general as well as to several negative regulators of DNA metabolism (Figure C.3). This negative regulation is achieved in part by chromatin remodeling, silencing, and heterochromatin formation. Thus, hsa-mir-130b may act to indirectly up-regulate DNA metabolism by down-regulating chromatin-based repressors.

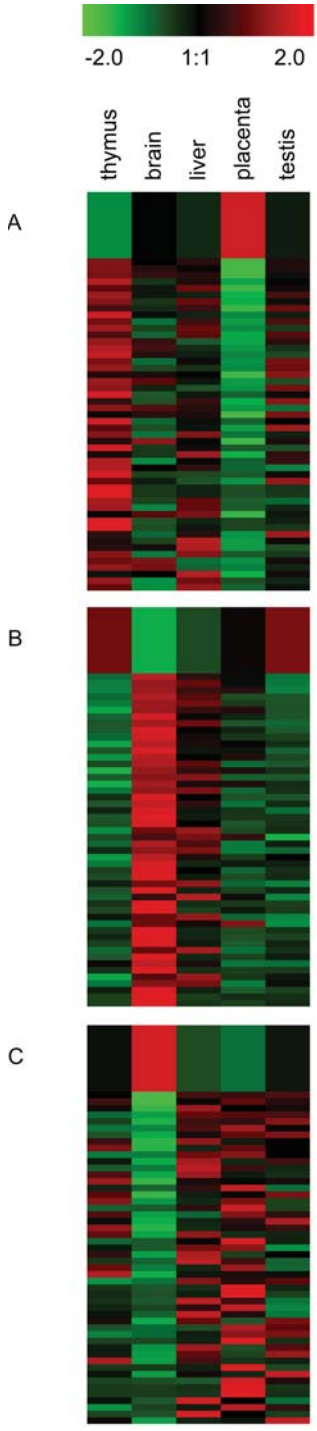

Figure 4.6: Anti-correlated expression patterns for TE-derived miRNAs and their targeted mRNAs. Results for three TE-derived miRNAs with expression data are shown: hsa-mir-130b (A), hsa-mir-28 (B) and hsa-mir-95 (C). The top row shows the relative miRNA expression across five human tissues, and the subsequent rows show relative expression levels for targeted mRNAs. The 50 most negative Pearson correlation coefficients (range r=-0.99 to -0.51; P=1.2e-10 to 1.3e-1) are shown for each plot.

Prediction of novel TE-derived miRNAs

The function of miRNAs, and of noncoding RNAs in general, is related to their secondary structure (MATTICK and MAKUNIN 2006). Selective constraint on such sequences often leads to compensatory mutations that maintain the base-pair interactions in the double-stranded regions of the structures, such as miRNA stem regions. Sequence alignments can be evaluated for the signal of conserved base-pair interactions as well as compensatory mutations to identify conserved, and thus presumably functionally relevant, secondary structural elements. Recent application of such techniques has led to the discovery of many novel putative regulatory RNA sequences (PEDERSEN *et al.* 2006; WASHIETL *et al.* 2005b). It has even been shown that orthologous regions that are not constrained at the level of primary sequence may nevertheless encode conserved secondary structural elements (TORARINSSON *et al.* 2006). Given the contribution of TEs to experimentally characterized miRNAs shown here and elsewhere (BORCHERT *et al.* 2006; PIRIYAPONGSA and JORDAN 2007; SMALHEISER and TORVIK 2005), we sought to evaluate human TE sequences for the ability to form hairpin structures along with the signals of conserved base pairs and compensatory mutations that indicate putatively functional secondary structures. This approach provides a way to predict further contributions of TEs to miRNAs.

Human genome TE sequences were evaluated for the potential to encode conserved secondary structures (PEDERSEN *et al.* 2006) that meet the criteria of miRNA genes (BENTWICH *et al.* 2005). This approach is conservative in the sense that it relies on sequence conservation and most of the experimentally characterized TE-derived miRNAs that we observe (37 of 55) are not evolutionarily conserved. Using this conservative

approach, we found 587 human TEs with the potential to encode conserved secondary structures (Table C.2); 4 of these sequences corresponded to previously known human miRNAs annotated in miRBase. Evaluation of these conserved secondary structures was used to identify 85 TE-derived sequences that meet the structural criteria of putative miRNA genes, and 70 of these sequences also show evidence of being expressed (Table 4.4). These 70 putative TE-derived miRNA sequences meet the previously defined biogenesis, conservation, and, at least in principle, expression criteria used for the identification of miRNA genes (AMBROS *et al.* 2003).

An example of a predicted TE-derived miRNA gene is shown in Figure 4.7. The MER135 sequence shown is a member of a family of recently characterized non-autonomous DNA-type elements, *i.e.*, MITEs, with ~500 copies in the human genome (JURKA 2006). Since MITEs have palindromic structures with terminal inverted repeats that flank short internal regions, their expression as RNA results in the formation of the kinds of hairpins seen for pre-miRNAs. Indeed, MITEs have previously been shown to contribute miRNA genes in the Arabidopsis and human genomes (METTE *et al.* 2002; PIRIYAPONGSA and JORDAN 2007).

Table 4.4: Predicted TE-derived miRNA genes

Name[a]	Coords[b]	TE[c]	Expression data[d]
3715_0_+_61	chr1:3131597-3131629(+)	MER121	EST/mRNA/KG/RS
15086_0_-_78	chr1:15041842-15041859(-)	HAL1	EST/mRNA/KG/RS
25288_0_-_83	chr1:23621848-23621877(-)	MIRb	EST/mRNA/KG/RS
30647_0_+_38	chr1:27752374-27752433(+)	MIRb	EST/mRNA/KG/RS
52664_0_-_50	chr1:44571346-44571464(-)	Eulor9A	EST/mRNA/KG/RS
67626_0_-_76	chr1:57127400-57127465(-)	Eulor1	EST/mRNA/KG/RS
85615_0_+_83	chr1:76474930-76474947(+)	MIRb	EST/mRNA/KG/RS
120809_0_+_79	chr1:111021701-111021719(+)	MIR	EST/mRNA
122080_0_-_62	chr1:112177611-112177631(-)	MIR	EST/mRNA/KG/RS
124780_0_-_66	chr1:114214379-114214407(-)	MIRb	EST/mRNA/KG/RS
154818_0_-_64	chr1:162825371-162825437(-)	MER135	EST/mRNA/KG/RS

Table 4.4 continued

Name[a]	Coords[b]	TE[c]	Expression data[d]
188052_1_-_92	chr1:198460508-198460590(-)	Eulor3	-
204532_0_-_104	chr1:211522027-211522054(-)	UCON31	EST
230542_0_-_67	chr1:244286075-244286098(-)	L1MB3	EST/mRNA/KG/RS
1231553_0_+_75	chr2:67238894-67239028(+)	Eulor4	EST/mRNA
1258257_0_+_85	chr2:104314401-104314489(+)	MER134	-
1361323_0_+_57	chr2:213067475-213067509(+)	Eulor5A	EST/mRNA/KG/RS
1573547_0_+_44	chr3:61643441-61643518(+)	MER126	EST/mRNA/KG/RS
1573643_0_+_95	chr3:61718341-61718381(+)	MER134	EST/mRNA/KG/RS
1620066_0_-_64	chr3:116298434-116298458(-)	Eulor1	EST/mRNA/KG/RS
1651767_0_+_52	chr3:146074810-146074873(+)	Eulor3	-
1668216_0_-_58	chr3:168436231-168436447(-)	MER126	-
1730972_0_-_56	chr4:46681709-46681733(-)	L1ME3B	EST/mRNA/KG/RS
1747758_0_-_63	chr4:74275595-74275629(-)	L1M5	EST/mRNA/KG/RS
1757379_0_+_70	chr4:85466757-85466855(+)	MER134	-
1827751_0_+_75	chr4:181988895-181988914(+)	MIRb	EST
1830405_0_+_49	chr4:183690755-183690850(+)	MER135	EST/mRNA/RS
1873731_0_+_53	chr5:58495675-58495729(+)	UCON9	EST/mRNA/KG/RS
1902777_0_+_53	chr5:90643387-90643420(+)	AmnSINE1_GG	EST/mRNA
1920501_0_+_72	chr5:113735156-113735173(+)	L2	EST/mRNA/KG/RS
1966281_0_+_83	chr5:156681824-156681841(+)	MIR3	EST/mRNA/KG/RS
1975838_0_-_80	chr5:165688874-165688944(-)	Eulor5A	-
1979031_0_+_61	chr5:167506770-167506888(+)	Eulor9A	EST/mRNA/RS
1987527_0_+_59	chr5:175727565-175727628(+)	L2	EST/mRNA/KG/RS
2000476_0_-_85	chr6:8499794-8499914(-)	Eulor6C	EST/mRNA
2031067_0_+_44	chr6:39048083-39048162(+)	Eulor5A	EST/mRNA/KG/RS
2075048_0_-_91	chr6:94484941-94484963(-)	ERVL-E	EST/mRNA
2115069.5_0_+_82	chr6:141179709-141179763(+)	Eulor5B	-
2165103_0_+_104	chr7:28447122-28447144(+)	MER121	EST/mRNA/KG/RS
2195049_0_+_117	chr7:73161289-73161306(+)	MIR3	EST/mRNA/KG/RS
2232211_0_+_45	chr7:113190696-113190791(+)	Eulor6B	-
2247695_1_+_65	chr7:129521966-129521985(+)	L1ME4a	EST/mRNA/KG/RS
2265159_0_+_85	chr7:146833245-146833271(+)	UCON4	EST/mRNA/KG/RS
2330918_0_-_108	chr8:79081399-79081462(-)	Eulor3	-
2344217_0_+_65	chr8:97188471-97188580(+)	MER135	EST
2348773_0_+_51	chr8:102229956-102230022(+)	Charlie9	-
2401146_0_-_96	chr9:16787222-16787246(-)	MIR	EST/mRNA/KG/RS
2421368_0_-_79	chr9:37811135-37811158(-)	L1MC4a	EST/mRNA/KG/RS
2426661_0_+_64	chr9:70297285-70297306(+)	MER91A	EST/KG/RS
2455634_0_-_64	chr9:105918396-105918420(-)	MER5A	EST/mRNA/KG/RS
2469999_0_+_79	chr9:118715772 -118715795(+)	UCON11	EST/mRNA/KG/RS
2500550_0_-_83	chrX:10899595-10899617(-)	L4	EST/mRNA/KG/RS
2519737_0_+_67	chrX:24557155-24557175(+)	L1ME4a	EST/mRNA/KG/RS
2598753_0_+_171	chrX:123865376-123865447(+)	Eulor11	EST/mRNA/KG/RS
2607024_0_-_68	chrX:131689852-131689873(-)	L1MB5	EST/mRNA/KG/RS
2625375_0_+_86	chrX:152562536-152562556(+)	L2	EST/mRNA/KG/RS
276291_0_+_66	chr10:62836157-62836220(+)	L1M5	-
285555_0_+_63	chr10:72980870-72980944(+)	MER125	EST/mRNA/KG/RS
334961_0_+_78	chr10:117579937-117579954(+)	L2	EST/mRNA/KG/RS
335779_0_+_54	chr10:118027456-118027512(+)	Eulor6D	EST
377681_0_+_96	chr11:19331037-19331062(+)	L3	mRNA/KG
425555_0_+_71	chr11:71985685-71985701(+)	MIR	EST/mRNA/KG/RS
438439_0_+_83	chr11:83316376-83316398(+)	L2	EST/mRNA/KG/RS
486187_0_+_68	chr11:130861130-130861151(+)	MIRb	EST/mRNA/KG/RS
487071_2_+_103	chr11:131453921-131453949(+)	MER122	EST/mRNA/KG/RS
492576_0_-_95	chr12:2125422-2125443(-)	MIRb	mRNA/KG/RS
533638.0_0_-_122	chr12:50492331-50492353(-)	MIRb	mRNA/KG
542148_0_-_83	chr12:55246557-55246574(-)	LTR37B	EST/mRNA/KG/RS

Table 4.4 continued

Name[a]	Coords[b]	TE[c]	Expression data[d]
551096_0_-_85	chr12:64538090-64538148(-)	Eulor5A	EST/mRNA/KG/RS
596947_0_+_93	chr12:115505370-115505426(+)	MER123	EST/mRNA
697653_0_+_69	chr14:33093444-33093479(+)	UCON11	EST/mRNA/KG/RS
700890_0_-_65	chr14:35855217-35855366(-)	Eulor6A	EST/mRNA/KG/RS
775713_0_+_77	chr15:25703141-25703162(+)	L1MCc	EST/mRNA/KG/RS
787092_0_-_65	chr15:35993736-35993832(-)	Eulor5A	-
896537_0_+_81	chr16:30749660-30749680(+)	MIR	EST
928869_0_+_74	chr16:70304015-70304037(+)	MIR3	EST/mRNA/KG/RS
976169_0_+_86	chr17:24040248-24040268(+)	L1ME4a	EST/mRNA/KG/RS
989909_0_+_100	chr17:34009010-34009024(+)	MIR3	EST/mRNA/KG/RS
1000039.8_0_+_109	chr17: 39468501-39468532(+)	L1MC4	EST/mRNA/KG/RS
1077028_0_-_58	chr18:33875730-33875789(-)	MIRb	-
1105916_0_-_78	chr18:71369451-71369514(-)	UCON11	-
1435354_0_-_79	chr20:44235903-44235921(-)	MIR	EST/mRNA/KG/RS
1443968_0_-_61	chr20:53838763-53838824(-)	UCON29	-
1466070_0_-_70	chr21:33853177-33853203(-)	L2	EST/mRNA
1496941_0_+_79	chr22: 35289947-35289989(+)	L1MC4	EST

[a]Name of the EvoFold locus from the hg18 UCSC Genome Browser annotation. The last field in the name corresponds to the EvoFold score.
[b]Genome coordinates and strand of the EvoFold locus
[c]Name of the co-located TE
[d]Source of the expression data for the locus: KG=UCSC Genome Browser Known gene annotation, RS=NCBI RefSeq annotation

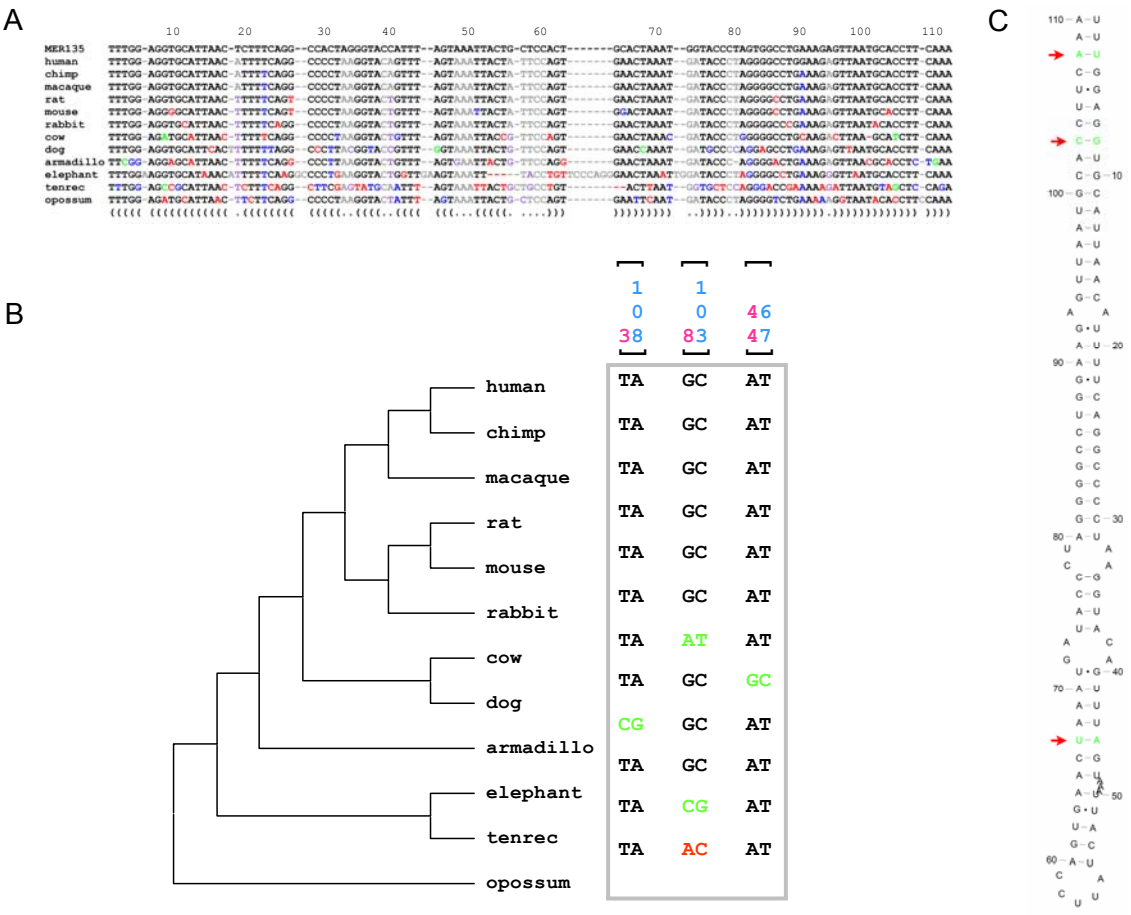

Figure 4.7: *Ab initio* prediction of a human TE-derived miRNA genes. (A) Multiple sequence alignment of the MER135 consensus sequence with the human genome sequence and orthologous genomic regions from 11 other vertebrate genomes. The predicted secondary structure is shown below the alignment with paired and unpaired positions indicated by '( )' and '.' characters respectively. Residues are colored according to the annotated secondary structure base pairs and their substitutions: gray=unpaired & no substitution, purple=unpaired & substitution, black=paired & no substitution, blue=paired & single substitution, green=paired & double substitution, red=not compatible with annotated pair. (B) Phylogenetic tree of the aligned species showing the double substitutions that maintain the secondary structure. Paired double substitutions are indicated with brackets and their positions in the alignment are shown. (C) Secondary structure of the predicted miRNA gene. Positions of the double substitutions are indicated with red arrows.

DISCUSSION

Abundance of TE-derived miRNAs

Noncoding regulatory RNAs, such as miRNAs, are a recently discovered class of genes, and the number of miRNA genes that exist among eukaryotic genomes is very much an open question (BEREZIKOV *et al.* 2006). Sustained efforts at high throughput characterization of miRNA genes, based on both experimental and computational approaches, continue to result in the discovery of many novel miRNAs (BENTWICH *et al.* 2005; CUMMINS *et al.* 2006). This can be appreciated by examining the release statistics of miRBase (ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/CURRENT/README). Plotting the number of miRNA gene entries against the miRBase release dates suggests that the number of known miRNA genes has experienced two distinct phases of linear increase, before and after the June 2005 release, and the current rate of increase in known miRNA genes is even greater than for the initial phase (Figure C.4).

For the most part, the miRBase data do not include substantial numbers of computationally predicted miRNA genes. The only computational predictions represented in miRBase are highly conserved sequences that are orthologous to experimentally characterized miRNAgenes in other species. Consideration of computationally identified miRNAs would suggest that miRNA gene numbers are substantially higher than currently appreciated. However, a number of computational methods for miRNA prediction do not consider TE-derived miRNAs (BENTWICH *et al.* 2005; LI *et al.* 2006; LINDOW and KROGH 2005; NAM *et al.* 2005). This is because, mainly for reasons of tractability, one of the first steps in computational analysis of eukaryotic genome sequences is the exclusion of repetitive DNA by RepeatMasking. TEs

will also tend to be excluded from predictions based solely on conservation between species because they are rapidly evolving and lineage-specific genomic elements. This is underscored by the fact that the set of TE-derived human miRNAs that we identify here is enriched for genes experimentally characterized in humans (93% for TE-derived *vs.* 81% for non TE-derived miRNAs; χ^2=4.76 *P*=0.03).

The factors described above that suggest the exclusion of TE-derived miRNAs led us to speculate as to how many more miRNA genes would be discovered if TE sequences were not eliminated from consideration *a priori*. To investigate this, we employed our own *ab initio* computational approach to try and predict TE-derived miRNA sequences. Application of this method to the human genome revealed 587 cases of human TE sequences that encode conserved RNA secondary structures, 85 of which are most likely to represent *bona fide* miRNA genes. Fifteen of the TE-derived miRNA genes that we predicted using this approach overlap with previous miRNA computational predictions (BEREZIKOV *et al.* 2005; PEDERSEN *et al.* 2006) as well as experimentally characterized miRNAs from miRBase.

Conservation of TE-derived miRNAs

Many miRNA genes are evolutionarily conserved and may have functional orthologs in multiple species. Indeed, sequence conservation is one of the criteria used to aid the computational discovery of miRNAs. While the TE-derived miRNA genes analyzed here are less conserved, on average, than non TE-derived miRNAs, there are a number of well-conserved miRNAs that evolved from TE sequences (Table 4.1). The majority of these conserved miRNAs are related to the ancient L2 and MIR TE families, and some of these sequences have been previously identified (SMALHEISER and TORVIK

2005). This is particularly interesting because numerous L2 and MIR sequences have been shown to be anomalously conserved between the human and mouse genomes (SILVA *et al.* 2003). Specifically, Silva et al. (2003) demonstrated that many L2 and MIR sequences found in orthologous human-mouse intergenic regions were present in the common ancestor of the two species and, following their divergence, evolved under strong selective constraint. From this, they reasoned that these selectively constrained sequences probably play some role related to gene regulation, although no specific functional role was ascribed to them. Here, we show that at least some of these conserved L2 and MIR fragments provide miRNA sequences with the potential to regulate numerous human genes.

As in the case of L2 and MIR (SILVA *et al.* 2003), comparative genomic approaches are used to infer functionally important genomic regions, particularly noncoding regions, by virtue of their high sequence conservation (ZHANG and GERSTEIN 2003). It is becoming increasingly apparent that a number of such highly conserved genomic sequences correspond to TEs (BEJERANO *et al.* 2006; KAMAL *et al.* 2006; NISHIHARA *et al.* 2006; XIE *et al.* 2006). While enhancer activity has been demonstrated for one of these conserved TEs (BEJERANO *et al.* 2006), for the most part, the specific function encoded by conserved TE sequences remains unknown. The collection of conserved TE sequences recently assembled by Repbase corresponds to <1% of all human genome TEs, but these sequences contribute >50% of all TE-encoded conserved secondary structures that we detected (Figure 4.2). Thus, our results suggest that many conserved TE sequences may encode miRNAs or perhaps other noncoding regulatory or structural RNAs.

Lineage-specific effects of TE-derived miRNAs

Most of the TE-derived miRNAs analyzed here are not evolutionarily conserved (Table 4.1). This is not surprising when you consider that TEs are the most lineage-specific and non-conserved elements found in eukaryotic genomes (LANDER *et al.* 2001). The over-representation of non-conserved sequences among TE-derived miRNAs is also consistent with previous work that has shown TE-derived cis-regulatory binding sites to be more divergent than non TE-derived cis sites (MARINO-RAMIREZ *et al.* 2005). From a practical perspective, this means that computational discovery methods that employ conservation as a criterion will necessarily overlook many TE-derived regulatory sequences. In terms of evolution, this means that the greatest differences between eukaryotic genomes will correspond to TE sequences. In this sense, TEs can be considered as drivers of genome diversification. This may be uninteresting if TEs serve only to replicate themselves and do not play any role for their host genomes as the selfish DNA theory of TEs holds (DOOLITTLE and SAPIENZA 1980; ORGEL and CRICK 1980). However, if some TEs are in fact functionally relevant to their hosts, as we have shown here for the case of TE-derived miRNAs, then their divergence may have important evolutionary implications. Indeed, TE-derived regulatory sequences may be particularly prone to contribute to regulatory differences among species that lead to lineage-specific phenotypes. This has been shown for the case of TE-derived regulatory sequences that are associated with high levels of expression divergence between humans and mice (MARINO-RAMIREZ and JORDAN 2006).

While most computational efforts to discover noncoding regulatory sequences have focused on conserved genomic elements, recent studies have begun to emphasize

rapidly evolving regions as well (POLLARD *et al.* 2006a; POLLARD *et al.* 2006b; PRABHAKAR *et al.* 2006). The rationale behind this is the notion that rapidly evolving regulatory regions may yield species-specific differences. An emphasis on the discovery of TE-derived regulatory sequences would complement current approaches to the discovery of rapidly evolving regulatory regions that are likely to contribute to the phenotypic divergence among species.

Genome defense and global gene regulatory mechanisms

Finally, we speculate that our results point to a connection between genome defense mechanisms necessitated by TEs and the emergence of global gene regulatory systems that may have allowed for the complex regulatory phenotypes characteristic of multicellular eukaryotes. TE insertions are highly deleterious and, as a consequence, a number of global gene silencing mechanisms, including methylation (YODER *et al.* 1997), imprinting (MCDONALD *et al.* 2005), and heterochromatin (LIPPMAN *et al.* 2004), may have evolved originally as TE defense mechanisms. siRNAs are also thought to have evolved as a defense mechanism against TEs (MATZKE *et al.* 2000; SLOTKIN *et al.* 2005; VASTENHOUW and PLASTERK 2004), and the results reported here and elsewhere (BORCHERT *et al.* 2006; PIRIYAPONGSA and JORDAN 2007; SMALHEISER and TORVIK 2005) indicate that miRNAs can emerge from TEs as well. More recently, an analogous TE defense mechanism based on small RNAs complementary to TEs in Drosophila has been reported (BRENNECKE *et al.* 2007). Apparently, different RNA interference systems may have evolved convergently on multiple occasions to help silence TEs. Later, these regulatory mechanisms could have been co-opted to exert controlling effects over thousands of host genes as is the case for miRNAs. The evolution of such complex gene

regulatory systems can be considered nonadaptive (LYNCH 2007) in the sense that they did not evolve by virtue of selection for the role that they play now. However, neither did these global regulatory mechanisms evolve passively since they were swept to fixation by selective pressure to defend against TEs. Therefore, the emergence of TE-related global regulatory systems, exemplified by RNA interference, can be considered to be exaptations (GOULD and VRBA 1982) driven by the internal mutational dynamics (STOLTZFUS 2006) of the genome.

CHAPTER 5

A FAMILY OF HUMAN MICRORNA GENES FROM MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS

ABSTRACT

While hundreds of novel microRNA (miRNA) genes have been discovered in the last few years alone, the origin and evolution of these non-coding regulatory sequences remain largely obscure. In this report, we demonstrate that members of a recently discovered family of human miRNA genes, hsa-mir-548, are derived from Made1 transposable elements. Made1 elements are short miniature inverted-repeat transposable elements (MITEs), which consist of two 37 base pair (bp) terminal inverted repeats that flank 6 bp of internal sequence. Thus, Made1 elements are nearly perfect palindromes, and when expressed as RNA they form highly stable hairpin loops. Apparently, these Made1-related structures are recognized by the RNA interference enzymatic machinery and processed to form 22 bp mature miRNA sequences. Consistent with their origin from MITEs, hsa-mir-548 genes are primate-specific and have many potential paralogs in the human genome. There are more than 3,500 putative hsa-mir-548 target genes; analysis of their expression profiles and functional affinities suggests cancer-related regulatory roles for hsa-mir-548. Taken together, the characteristics of Made1 elements, and MITEs in general, point to a specific mechanism for the generation of numerous small regulatory RNAs and target sites throughout the genome. The evolutionary lineage-specific nature of MITEs could also provide for the generation of novel regulatory phenotypes related to

species diversification. Finally, we propose that MITEs may represent an evolutionary link between siRNAs and miRNAs.

INTRODUCTION

Numerous human genome transcripts lack protein coding capacity, and these non-coding RNA (ncRNAs) perform a variety of structural, enzymatic and regulatory functions (MATTICK and MAKUNIN 2006). MicroRNAs (miRNAs) are a class of short, ~22 nt ncRNA that function as post-transcriptional regulators of gene expression (AMBROS 2004). Mature miRNAs are processed from longer RNA sequences that form local stem-loop (hairpin) structures (BARTEL 2004). The first step of the miRNA biogenesis pathway occurs in the nucleus where the RNase III enzyme Drosha cleaves both strands of the so-called pri-miRNA at the base of the stem. This yields a ~70–90 bp pre-miRNA hairpin that is exported to the cytoplasm where it is further processed by Dicer, another RNase III endonuclease. Dicer recognizes the double stranded portion of the RNA close the base of the pre-miRNA stem and cleaves both strands of the duplex in two places. This reaction cuts off the loop portion of the molecule as well as the terminal part of the stem leaving a short duplex that consists of the mature miRNA and a complementary miRNA* sequence that is rapidly degraded. Once liberated in this way, the mature miRNA sequence binds to partially complementary target sites in the 3' untranslated regions (UTRs) of messenger RNAs (mRNAs) and regulates expression through a process of mRNA degradation and/or translational repression (BARTEL 2004).

miRNAs were only recently discovered (LEE *et al.* 1993), and details regarding their origin and evolution have yet to be fully worked out. Since their original discovery, miRNAs have been detected in all metazoa surveyed for their presence (BARTEL 2004).

However, the full extent of miRNA genes in any particular genome is unknown, and a number of studies aimed at the detection of novel miRNA genes have been conducted to address this issue. Bioinformatic miRNA discovery relies primarily on the sequence conservation of miRNA genes and secondary structure of the pre-miRNAs (BENTWICH *et al.* 2005), while experimental efforts consist of forward (LEE *et al.* 1993) and reverse (CHEN *et al.* 2004) genetic studies as well as efforts to clone short mature miRNA sequences (LAGOS-QUINTANA *et al.* 2001; LAU *et al.* 2001; LEE and AMBROS 2001). Cloning mature miRNA sequences is technically challenging given their small size and associated instability. Thus, direct miRNA cloning is not well suited to large scale discovery efforts and may have already reached the point of diminishing returns (LAGOS-QUINTANA *et al.* 2001). A recently published report described a novel high-throughput miRNA cloning technique aimed at increasing the efficiency of miRNA discovery (CUMMINS *et al.* 2006). This technique is based on the serial analysis of gene expression (SAGE) and takes advantage of well established protocols tailored to small RNA sequences. Application of this SAGE-based approach to human transcripts confirmed the presence of numerous miRNA genes that had been detected previously through computational and/or experimental surveys and also yielded more than 100 novel miRNA sequences (CUMMINS *et al.* 2006). Including these new data, version 8.2 of miRBase, the online microRNA database (GRIFFITHS-JONES *et al.* 2006), reports 462 human miRNA genes. The importance of miRNAs for human gene regulation is underscored by target site predictions (ENRIGHT *et al.* 2003), which reveal that these human miRNAs have the potential to regulate thousands of human genes.

miRNAs are closely related to another class of ncRNA, known as small interfering RNAs (siRNA), in terms of both biogenesis and regulatory function (AMBROS *et al.* 2003; BARTEL 2004). The mature biologically active forms of siRNA and miRNA are both processed from double stranded RNA (dsRNA) by Dicer. However, siRNAs are generated from long dsRNA precursors, which can be either endogenous or exogenous transcripts, whereas mature miRNAs are processed from shorter endogenous transcripts that form local hairpin structures. Numerous siRNA molecules are processed from both strands of the long dsRNA precursor, whereas a single mature miRNA sequence is generated from only one strand of the pre-miRNA hairpin. While miRNAs can act through translational repression of their targets, they may also cause mRNA degradation of their target genes in the same way that siRNAs do (HUTVAGNER and ZAMORE 2002a; LLAVE *et al.* 2002; YEKTA *et al.* 2004; ZENG *et al.* 2003).

One previously recognized distinction between these two classes of regulatory RNA is the fact that miRNAs are generally found in unique genomic loci, such as intergenic regions (BARTEL 2004), while siRNAs originate from within already characterized sequences such as genes and transposable elements (TEs) (MATZKE *et al.* 2000; SLOTKIN *et al.* 2005; VASTENHOUW and PLASTERK 2004). However, a recent report indicated that a number of mammalian miRNAs, including six human miRNAs, are in fact derived from TEs (SMALHEISER and TORVIK 2005). The abundance and repetitive nature of TE sequences could provide a natural mechanism for the generation of multiple miRNA genes, along with homologous target sites, dispersed throughout the human genome. TEs may also provide an evolutionary connection between siRNAs and miRNAs. In light of these possibilities, we sought to investigate the relationship between

human miRNAs and TEs by evaluating whether there exist families of related (paralogous) miRNA genes that are derived from TE sequences. We compared the genomic locations of experimentally characterized human miRNA genes to the annotated human TE sequences and discovered a set of closely related miRNA genes derived from a family of miniature inverted repeat transposable elements (MITEs). The palindromic sequence structure of MITEs, considered together with their insertion into transcriptionally active regions of the human genome, suggests a specific mechanism by which these kinds of elements could give rise to emergent mature miRNAs.

METHODS

TE-miRNA sequence analysis

The UCSC Genome and Table Browsers (KAROLCHIK *et al.* 2003; KAROLCHIK *et al.* 2004) were used to analyze the March 2006 human genome reference sequence (http://www.genome.ucsc.edu/cgi-bin/hgGateway?org = Human&db = hg18). This sequence is referred to as the hg18 assembly on the UCSC Genome Bioinformatics website and corresponds to the human genome build 36.1 assembled by the National Center for Biotechnology Information (NCBI). The Table Browser was used to search genome-wide for co-located TE and miRNA gene sequences, and the Genome Browser was used to visualize the results on a case-by-case basis. The genome locations and identities of human TE sequences were taken from annotation generated by the RepeatMasker program (http://www.repeatmasker.org) (SMIT *et al.* 1996-2004). The genome locations and identities of experimentally characterized human miRNA gene sequences were taken from release 8.2 of the miRBase sequence database (http://microrna.sanger.ac.uk/sequences/) (GRIFFITHS-JONES *et al.* 2006). Evolutionary

conservation between human Made1-derived miRNA gene sequences and six mammalian genomes – chimp, rhesus, mouse, rat, dog and cow – was assessed based on the Alignment Net track of the UCSC Genome browser, which shows the best pairwise between-genome alignments corresponding to orthologous regions (KENT *et al.* 2003).

The sequences of Made1-derived miRNAs were compared to the human genome sequence using the BLAT program (KENT 2002). Homologous genomic sequences were counted as statistically significant hits that matched ≥80% of the length of the query miRNA sequence and were confined to a local genomic region no longer than 120% of the query length (*i.e.* long genomic insertions were not counted). Made1 and hsa-mir-548 sequences were aligned to each other using the program ClustalW (THOMPSON *et al.* 1994). NCBI's BLASTN program (http://www.ncbi.nlm.nih.gov/BLAST/) (ALTSCHUL *et al.* 1997) was used to search the Expressed Sequence Tags Database (dbEST) (BOGUSKI *et al.* 1993) for expressed human MITE sequences. Human genomic expression data from Affymetrix tiling GeneChips (CHENG *et al.* 2005), represented in the UCSC Genome Browser, were evaluated in order to identify transcriptionally active regions of the human genome. RNA sequences were folded using the Mfold (ZUKER 2003) web server (http://www.bioinfo.rpi.edu/applications/mfold/rna/form1.cgi).

Regulatory analysis

Putative miRNA target sites were taken from the miRBase Targets website (http://microrna.sanger.ac.uk/targets/v3/), which uses a modified implementation of the miRanda algorithm (ENRIGHT *et al.* 2003). 3' UTRs of Ensembl genes were also searched for Made1-derived target sites. In this case, the same approach used by the current miRBase implementation of miRanda for annotating 3' UTRs was employed.

Specifically, if there is no hexamer of 'A' residues in the last 30 bp of the 3' UTR, the sequence is extended 2,000 bp. The random expectation for the number of target genes identified by both methods was calculated by taking their joint probability multiplied by the total number of human genes (n = 23,269 from Ensembl version 41). The joint probability was calculated by multiplying the relative human genome frequencies of each target set. The difference between the expected and observed number of target genes identified using both methods was calculated using the binomial distribution.

Comparative genomic sequence data from the UCSC genome browser were used to analyze the relative evolutionary conservation levels for predicted hsa-mir-548 target sites. Position-specific conservation scores were derived from multiple whole genome sequence alignments between the human and 16 other vertebrate genomes (KENT *et al.* 2003; ZUKER 2003). The scores correspond to the posterior probability that a human genome site is conserved as determined by the phastCons program (SIEPEL *et al.* 2005), and position-specific scores were averaged across target sites.

Human gene expression patterns across 79 tissues were taken from the Novartis Research Foundation's Symatlas (SU *et al.* 2004). Relative expression profiles for genes with hsa-mir-548 target sites were computed for each gene by dividing the gene's tissue-specific expression (signal intensity) values by the gene's median expression value over all 79 tissues and then $\log_2$ normalizing the resulting ratios. The program Genesis (STURN *et al.* 2002) was used to visualize the relative expression profiles, to group related expression profiles with k-means clustering and to group tissues with hierarchical clustering.

Clusters of coexpressed genes were analyzed with the program GOTree Machine (GOTM) (ZHANG *et al.* 2004) to look for over-represented Gene Ontology (ASHBURNER *et al.* 2000) functional annotations. To do this, genes in each cluster were annotated with their biological process GO terms. The frequencies of these terms were then compared to their expected frequencies based on their occurrences in the human genome, and statistically over-represented terms were identified using the hypergeometric test. Statistically over-represented terms were then mapped to the GO directed acyclic graph.

RESULTS AND DISCUSSION

A TE-derived miRNA gene family

When we compared the genomic locations of experimentally characterized human miRNA sequences stored in miRBase (GRIFFITHS-JONES *et al.* 2006) to the locations of human TEs characterized by the program RepeatMasker (SMIT *et al.* 1996-2004), we found that seven closely related miRNA genes (hsa-mir-548) were co-located with dispersed members of a single family of TEs known as Made1 (Table 5.1). These hsa-mir-548 miRNA genes were recently characterized by mapping mature cloned miRNA sequences to the human genome sequence (CUMMINS *et al.* 2006). The hsa-mir-548 mature miRNAs meet both the expression and biogenesis criteria that were articulated to ensure the accurate identification of miRNAs and the distinction between miRNAs and siRNAs (AMBROS *et al.* 2003). In particular, the mature hsa-mir-548 miRNAs are all 22 nt in length, they were identified from a cDNA library made of size fractionated RNA and they map precisely to genomic regions that are predicted to form local hairpin structures.

Inspection of the multiple sequence alignment of a full length Made1 sequence with all seven hsa-mir-548 miRNAs provides clear evidence that the miRNAs are in fact derived from the Made1 elements (Figure 5.1). Individual hsa-mir-548 sequences were queried against the human genome sequence to search for duplicates. Each hsa-mir-548 gene showed significant similarity to numerous genomic regions (Table 5.1), suggesting the possibility that this miRNA gene family may include many as yet uncharacterized members.

Table 5.1: Made1-derived miRNA genes in the human genome

Name[a]	Accn[b]	Chr[c]	Start[c]	Stop[c]	Str[c]	Duplicates[d]
hsa-mir-548a-1	MI0003593	6	18679994	18680090	+	24
hsa-mir-548a-2	MI0003598	6	135601991	135602087	+	81
hsa-mir-548a-3	MI0003612	8	105565773	105565869	-	82
hsa-mir-548b	MI0003596	6	119431911	119432007	-	23
hsa-mir-548c	MI0003630	12	63302556	63302652	+	124
hsa-mir-548d-1	MI0003668	8	124429455	124429551	-	71
hsa-mir-548d-2	MI0003671	17	62898067	62898163	-	145

[a]miRNA gene name
[b]miRBase accession number
[c]Human genome chromosome coordinates and strand information
[d]Duplicate sequences taken as the number of statistically significant human genome BLAT hits that also pass the match length criteria described in the Methods section

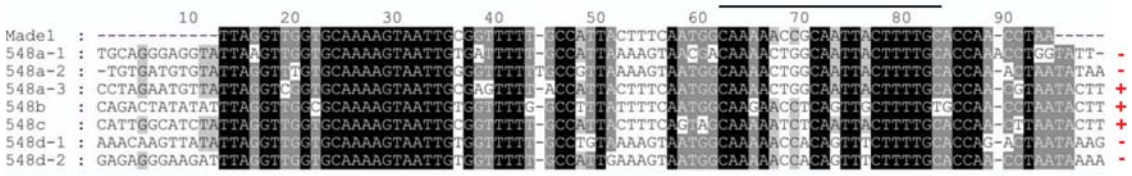

Figure 5.1: Multiple sequence alignment of Made1 and hsa-mir-548 genes. The location of the mature miRNA sequence is indicated by the bar over the alignment. The strand of the Made1 element (+/-) from which the miRNA genes are derived is shown to the right of the alignment.

Made1 elements were independently characterized by several groups as non-autonomous derivatives of the human mariner-like transposable element (TE) Hsmar1 (MORGAN 1995; OOSUMI *et al.* 1995; SMIT and RIGGS 1996). Hsmar1 elements are DNA-type TEs, approximately 1,300 bp in length, which possess a transposase-encoding open reading frame flanked by terminal inverted repeat (TIR) sequences (Figure 5.2A) (ROBERTSON and ZUMPANO 1997). Related Made1 elements are only 80 bp long with two 37 bp TIRs and a 6 bp intervening region (Figure 5.2B). In this sense, Made1 sequences are palindromes, and if they were to be transcribed, they would form highly stable hairpin loops reminiscent of the pre-miRNA structures that are processed to form mature miRNAs (Figure 5.2C).

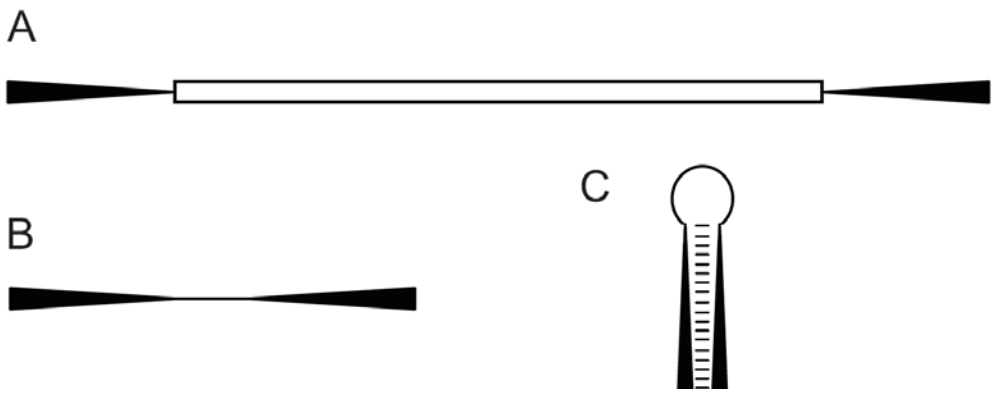

Figure 5.2: Schematic illustrating the relationship between Hsmar DNA-type TEs (A), Made1 MITEs (B) and hairpins (C) of the kind recognized by the miRNA enzymatic processing machinery. (A) A full length DNA-type element with terminal inverted repeats (TIRs) flanking an open reading frame (ORF) is shown. (B) Non-autonomous MITE derivative of a full length DNA-type element, containing TIRs but no internal ORF. (C) Predicted hairpin structure that would be formed by base-pair interactions of the MITE TIRs.

The formation of TIR-based dsRNA hairpins from Made1 would require the generation of full-length (or nearly so) element transcripts. The human expressed sequence tag (EST) database was searched using BLASTN (ALTSCHUL *et al.* 1997), with a full-length Made1 query sequence, to test for this. We found 141 human ESTs that showed >80% sequence similarity to the Made1 query sequence over >80% of the length of the element (Table D.1). Furthermore, the EST analysis indicates that Made1 sequences are widely expressed in a variety of tissue-types, providing ample opportunity for the formation of mature miRNAs.

Interestingly, Made1 transcripts destined to become hsa-mir-548 miRNAs are generated from both strands of the element (Figure 5.1). Because the element sequences are palindromes, transcripts produced in either orientation (+/-) would yield local hairpin structures. Indeed, the only difference between strand-specific transcripts is seen for the intervening 6 bp sequence that forms the loop in the structure (positions 51–56 in Figure 5.1). This suggests that Made1 expression may result from read-through transcripts promoted from adjacent genomic positions, as opposed to a strand specific promoter encoded by the element itself. Consistent with this notion, we found that a number of Made1 homologous ESTs include substantial upstream and downstream sequences (Table D.1).

Therefore, we propose a model whereby Made1 insertions into transcriptionally active genomic regions would yield viable pri-miRNA structures that would be processed into mature miRNA sequences by the RNA interference enzymatic machinery. An example of such a scenario can be seen for the human EST corresponding to the Genbank accession BU608159. This 754 base pair (bp) EST maps to chromosome 13 at positions

24,718,360–24,719,104; it includes a nearly full length Made1 element as well as 325 bp of 5' flanking DNA and 353 bp of sequence 3' to the element. Visualization of genomic expression data, generated with human genome tilling arrays (CHENG *et al.* 2005), shows that this particular Made1 is inserted into an intergenic region of the genome that is transcriptionally active (Figure 5.3). In this case, the entire Made1 element is transcribed as a read-through initiated from an adjacent genomic position. When the RNA structure of the EST, which includes the Made1 element along with expressed genomic flanking regions, is evaluated using the program Mfold, the Made1 region can be seen to form the most stable stem-loop structural element in the RNA (Figure 5.4A). The tight hairpin formed by the element is similar to the structures processed by Drosha and Dicer, and the location of the mature miRNA sequence, in the stem close to the 3' end of the structure, is consistent with the mode of cleavage thought to be employed by the Dicer (Figure 5.4B).

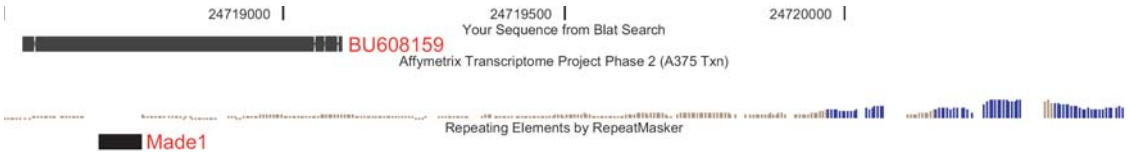

Figure 5.3: Made1 insertion in a transcriptionally active region of the human genome. The Made1 element shown is expressed by read-through from an adjacent promoter position in the genome. The EST BU608159 consists of the Made1 element along with 678 bp of flanking DNA.

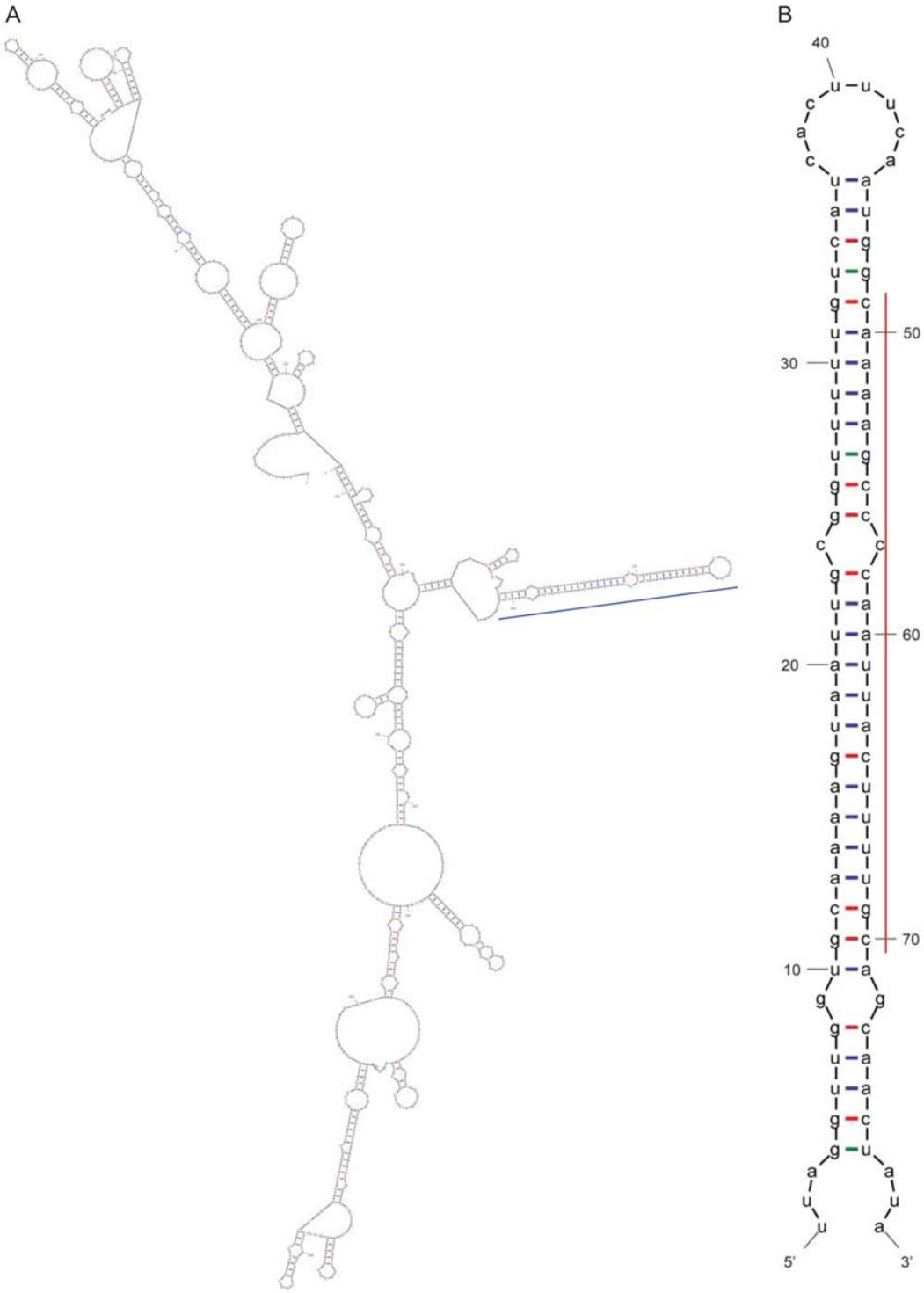

Figure 5.4: RNA secondary structures of the entire BU608159 EST (A) and the Made1 element contained within this transcript (B). The Made1 hairpin region of the BU608159 structure is indicated with a blue bar (A), and the location of the mature miRNA sequence is shown with the red bar (B).

Regulatory effects of hsa-mir-548

Mature miRNA sequences associate with the RNA-induced silencing complex (RISC), which facilitates their regulatory interactions with target mRNAs (BARTEL 2004). miRNAs wield specific regulatory effects on gene expression through physical interactions with partially complementary sequences in the 3' untranslated regions (UTRs) of their target genes' transcripts. We sought to characterize the potential regulatory and functional effects of hsa-mir-548 miRNAs by analyzing the genes that they are predicted to target.

Putative hsa-mir-548 target sites were identified using two methods: i-by the modified miRanda algorithm implemented in miRBase and ii-by searching 3' UTRs for Made1 sequences that are complementary to the mature hsa-mir-548 miRNAs. According to the miRBase predictions, the seven hsa-mir-548 genes have 3,527 potential target genes. Made1 related targets, on the other hand, are found in only 179 genes. This was slightly surprising given that there are 7,850 annotated Made1 sequences in the human genome. When the search for Made1-derived target sites was extended to entire transcripts, only one additional target was found in a 5' UTR. Apparently, Made1 sequences avoid protein coding gene exon regions and thus are poorly represented among potential hsa-mir-548 target sites. Furthermore, the intersection of the target gene sets derived from the miRBase versus Made1 consists of a mere 29 genes, and this figure is only slightly higher than the random expectation of 27 shared targets (P=0.07 binomial distribution). That both target prediction methods detect such a small number of the same targets can be attributed to the fact that Made1 targets are likely to be avoided by the miRanda based approach due to its criterion of sequence conservation and the fact that

Made1 is an evolutionarily young TE family. Indeed, when the sequence conservation levels of target sites identified by the two methods were compared, Made1 related targets were found to be significantly less conserved, on average, than miRanda predicted targets (conservation scores: Made1 targets=0.0826±0.017 miRanda targets=0.3196±0.007; t=11.27 P=5.7e-29 Student's t-test).

The potential functional relevance of genes with Made1-derived target sites was evaluated by considering their Gene Ontology (GO) biological process annotations and looking for over-represented functional categories. This procedure identified seven over-represented GO biological process categories that include a total of 11 genes (Table D.2). The relationships among the over-represented GO functional categories in the Made1 target gene set can be visualized on the GO directed acyclic graph (Figure 5.5). This set includes genes with functional roles in cell proliferation, mitosis and apoptosis, all categories that are related to cancer. The hsa-mir-548 genes were characterized by virtue of their expression in colorectal cancer cell lines and tissue samples (CUMMINS *et al.* 2006). If hsa-mir-548 expression is up-regulated in colorectal cancer tissue, it may lead to the repression of genes that normally control cellular proliferation. Consistent with this scenario, several of the genes that correspond to over-represented functional categories were found to be down regulated in colorectal cancer tissue (Table D.2). These include genes encoding a cell division cycle protein (ENSG00000004897), a C epsilon type protein kinase (ENSG00000171132) and a centromere/kinetochore protein (ENSG00000086827).

As mentioned previously, the paucity of Made1 related target sites was somewhat unexpected. Nevertheless, the identification of numerous non-Made1 related target sites

is interesting in the sense that it suggests that TE-derived miRNAs may be able to regulate host genes that do not have any related TE sequences. There are two models to explain the repressive effects that miRNAs exert on target gene expression: i-translational repression and ii-mRNA degradation (BABAK *et al.* 2004; YEKTA *et al.* 2004; ZENG *et al.* 2003). Recently, anti-correlated expression patterns between miRNA sequences and their target mRNAs have provided evidence in favor of the mRNA degradation model (HUANG *et al.* 2006). We sought to further evaluate the potential mRNA degradation-based regulatory effects of the hsa-mir-548 miRNAs by searching for down regulation of putative target genes in tissue samples similar to the colorectal samples from which they were cloned (CUMMINS *et al.* 2006). Consideration of target gene relative expression levels can also be used to help validate target site predictions, which are prone to false positives.

Gene expression profiles for potential hsa-mir-548 targets were taken from the Novartis Research Foundation's Symatlas (SU *et al.* 2004). For the miRBase set, a total of 2,045 target genes were found with corresponding Symatlas expression data across 79 human tissues. The expression data were median and log normalized to yield relative tissue-specific gene expression profiles, and these profiles were separated into 20 co-expressed groups of genes using k-means clustering. Three of these clusters – 12, 15 and 20 –showed marked down-regulation of the colorectal adenocarcinoma sample (Figure 5.6). Interestingly, the genes found in these same clusters tended to be down-regulated in all five of the other cancer-related samples in the data set (Figure 5.7). This suggests the possibility that hsa-mir-548 miRNA genes may play some global role related to the regulation of gene expression in cancer. Indeed, hierarchical clustering of the tissue-

samples based on the gene expression data unites all of the cancer samples into a single group to the exclusion of all normal tissues (Figure D.1); however, the colorectal adenocarcinoma sample is the outlier of this group (Figure 5.8). When the $\log_2$ median expression ratios were averaged for all genes with putative hsa-mir-548 target sites, the colorectal sample had the lowest relative expression level (q=9.72, v=12738, k=6, P<0.001 Tukey test; Figure 5.8). This finding is consistent with the fact that the hsa-mir-548 genes were isolated from colorectal cancer samples, and points to an additional more specific role for these genes in colorectal cancer related gene regulation. The functional affinities of the genes in the three down regulated clusters were assessed using the same GO-based approach as for the set of genes with Made1 target sites. There are 29 GO biological process categories, encompassing 104 genes, which contain an over-representation of genes from these clusters (Table D.3). These include genes involved in cell adhesion, cell signalling and signal transduction. The positions of these categories on the GO biological process DAG can be seen in Figure D.2.

We also compared putative hsa-mir-548 target genes to a recently published collection of genes that were indicated as being involved in colorectal cancer by microarray expression profiling (SHIH *et al.* 2005). We found 22 examples of putative hsa-mir-548 target genes that were previously found to be related to colorectal cancer based on down-regulation in six separate microarray studies (Table D.4). These include a number of genes encoding various immune cell receptors as well as transcription factors and tumor necrosis factors. The apparent connection between cancer and the immune system in our dataset is supported by the similar down-regulated expression patterns seen for hsa-mir-548 target genes among the cancer and immune tissue samples (Figure 5.7).

However, a number of genes previously implicated in colorectal cancer etiology by virtue of up-regulation in previous studies were also found to have predicted hsa-mir-548 target sites. These cases may represent false positive target site predictions or could point to instances where hsa-mir-548 miRNAs act through translational repression and thus do not repress mRNA expression levels.

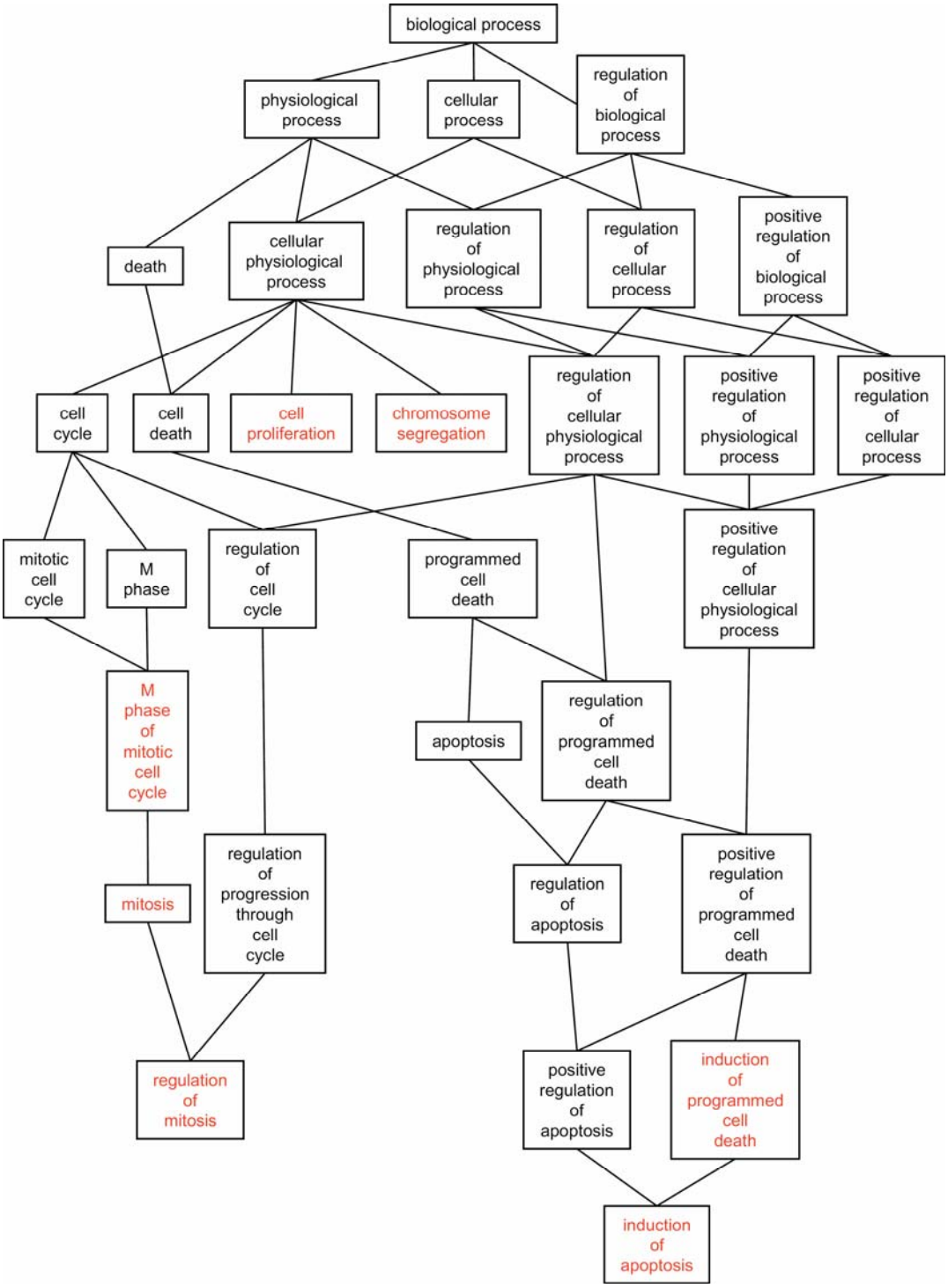

Figure 5.5: GO biological process terms over-represented among the set of genes with Made1-derived hsa-mir-548 target sites. The portion of the directed acyclic graph (DAG) containing all paths from the root biological process term to the over-represented functional category terms is shown. Over-represented functional categories are indicated in red.

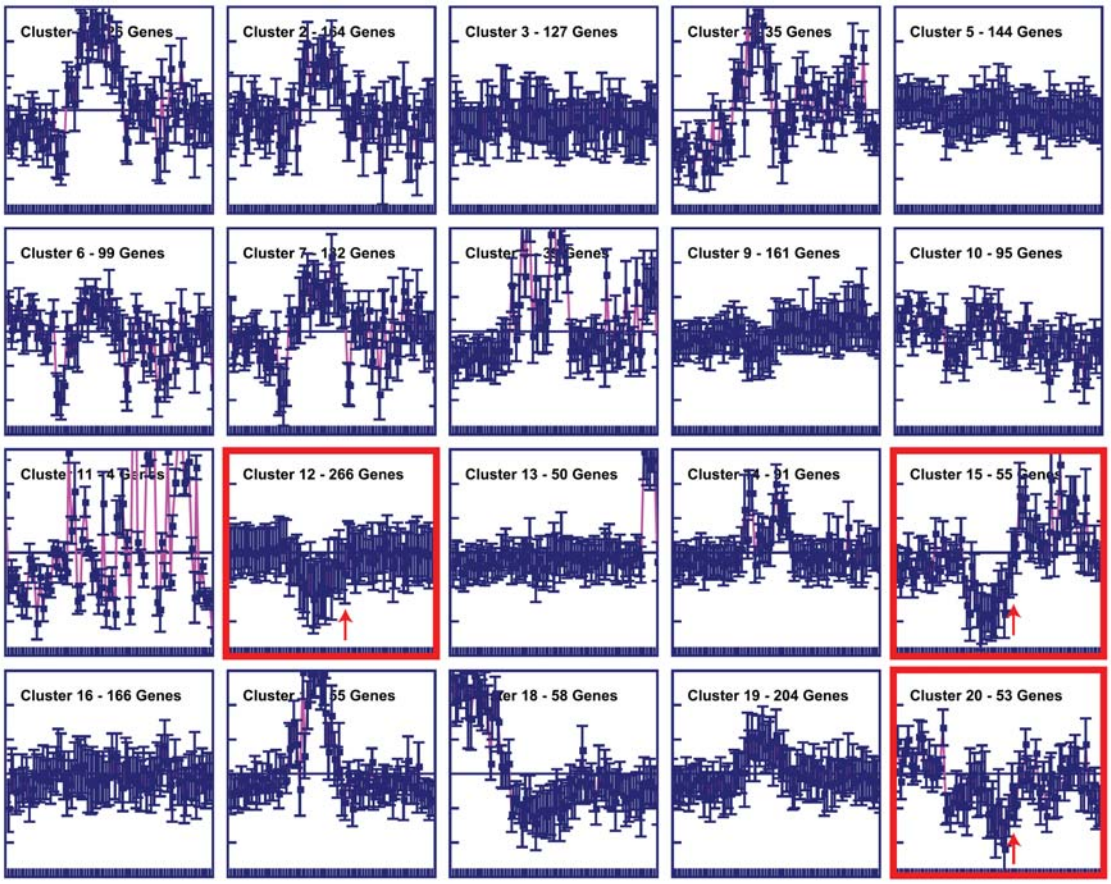

Figure 5.6: Coexpressed clusters of putative hsa-mir-548 target genes. Centroid views with average tissue-specific expression values are shown for all 20 clusters. Clusters containing genes down-regulated in the colorectal adenocarcinoma sample are shown in red and arrows indicate the colorectal sample.

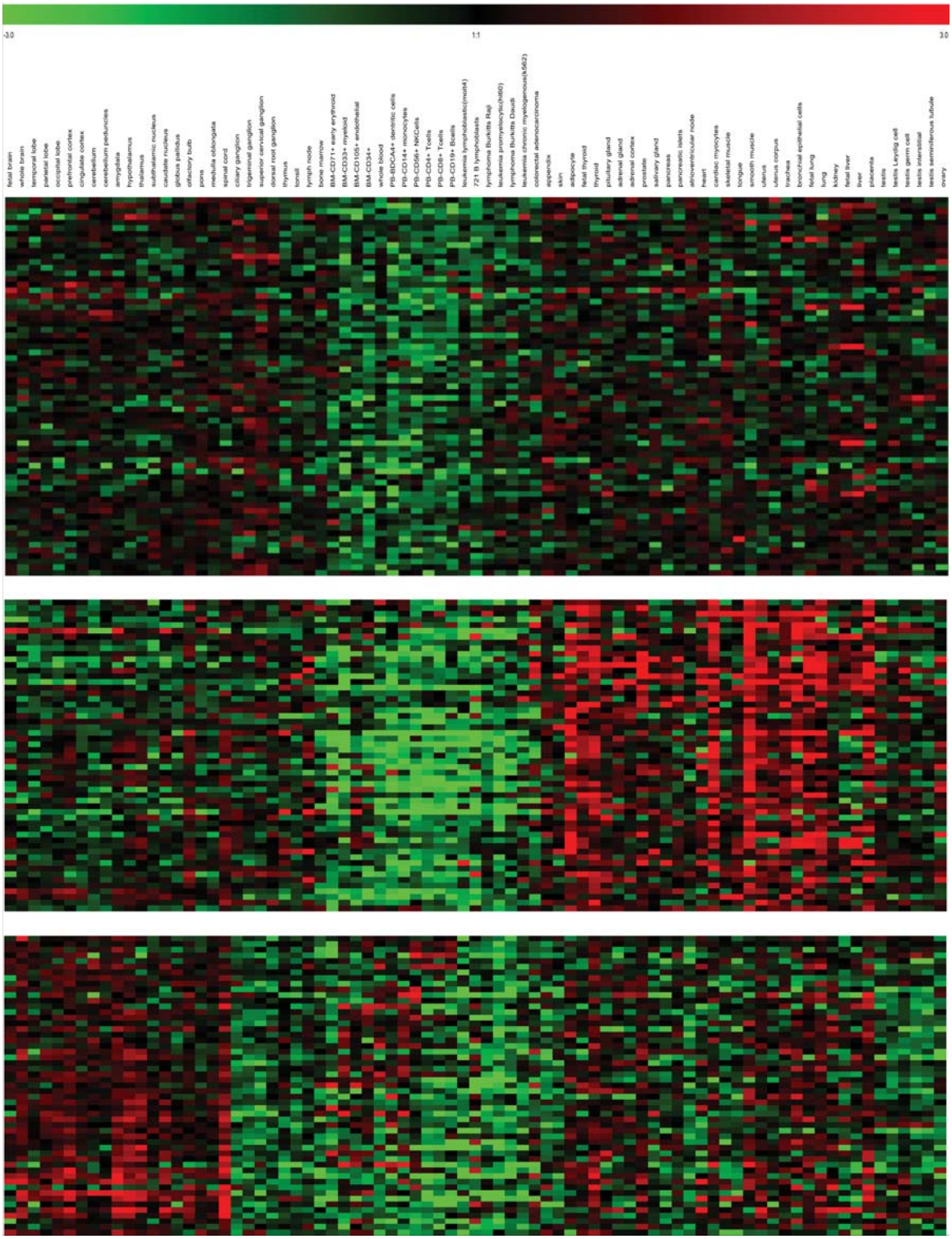

Figure 5.7: Representative gene expression profiles for putative hsa-mir-548 target genes from three coexpressed clusters (12, 15 and 20 in Figure 5.6). Expression profiles are median centered and $\log_2$ normalized, and the $\log_2$ ratio color scale is shown above the plot. Overexpressed genes are shown in red and underexpressed genes are shown in green.

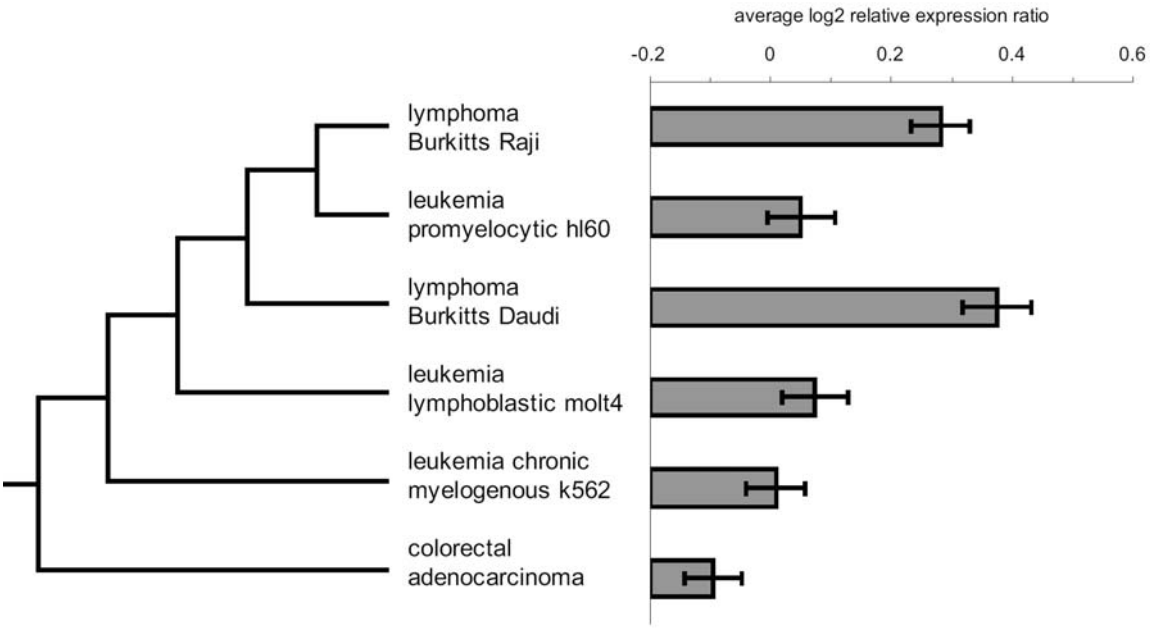

Figure 5.8: Relationships and average relative expression levels among the cancer tissues samples from the Novartis Symatlas microarray dataset. A dendogram relating the cancer samples based on similarities (differences) among relative expression levels is shown along with the average relative expression levels for all genes with hsa-mir-548 target sites in each of the cancer samples.

CONCLUSIONS

We report here a human miRNA gene family derived from TEs. The palindromic structure of the Made1 elements from which the hsa-mir-548 miRNA genes originated, together with their insertion into transcriptionally active genomic regions, points to a specific mechanism by which these sequences can be recognized and processed by the enzymatic machinery that yields mature miRNA sequences. In addition, the dispersed repetitive nature of TE sequences provides for the emergence of multiple novel miRNA genes as well as numerous homologous target sites throughout the genome.

TEs also tend to be among the most lineage-specific, *i.e.* recently evolved, sequences in the human genome (LANDER *et al.* 2001). Made1 elements emerged along

the primate evolutionary lineage, and orthologous hsa-mir-548 sequences are confined to the human, chimpanzee and rhesus macaque genome sequences (Figure D.3). While many miRNA genes are conserved across more distantly related species, a recent analysis of the human genome detected numerous putative miRNAs that are not evolutionarily conserved (BENTWICH *et al.* 2005). TEs, such as Madel, represent a natural source of such lineage-specific miRNAs, which could in turn be responsible for regulatory phenotypes that contribute to evolutionary diversification between species. The relatively low conservation of Made1-derived target sites is also consistent with this lineage-specific mode of evolution.

MITEs are widely distributed among eukaryotes (FESCHOTTE *et al.* 2002a) and could provide for the emergence of regulatory RNAs, such as miRNAs, siRNAs or other small non-coding RNAs, in many different genomic contexts. For instance, MITEs are particularly prevalent in plants where they were first discovered (BUREAU and WESSLER 1992); the rice genome alone contains ~90,000 MITEs (JIANG *et al.* 2004). A striking feature of plant MITEs is their apparent preference for insertion in gene regions (MAO *et al.* 2000; ZHANG *et al.* 2000). Accordingly, many thousands of plant MITEs must be expressed along with the gene sequences in which they are inserted. This would provide ample opportunities for the processing of MITE hairpins by RNA interference enzymatic machinery, which is known to play a particularly important role in plant gene regulation (MATZKE and MATZKE 2004).

Finally, we would like to propose that MITEs, such as Made1, may represent an evolutionary intermediate between siRNAs and miRNAs. A number of epigenetic gene silencing mechanisms, such as cytosine methylation (YODER *et al.* 1997), genomic

imprinting (MCDONALD *et al.* 2005) and heterochromatin (LIPPMAN *et al.* 2004) are thought to have evolved as defense mechanisms against transposition. Subsequently, these TE silencing mechanisms were co-opted as global regulators to control the expression patterns of host genes. This may have led to the increase in regulatory and phenotypic complexity seen among members of the eukaryotic crown group. In a similar way, RNA interference by siRNAs is considered to have evolved to silence TEs (MATZKE *et al.* 2000; VASTENHOUW and PLASTERK 2004). Consistent with this model, there are a number of cases of siRNAs that originate from TEs in different species (ARAVIN *et al.* 2003; HAMILTON *et al.* 2002; LIPPMAN *et al.* 2003; ZILBERMAN *et al.* 2003). Perhaps the best characterized example of this is the Muk TE repressor in maize (SLOTKIN *et al.* 2005). Muk is an effective silencer of the MuDR DNA-type TE, and the Muk locus consists of an inverted duplication of a partially deleted MuDR element. When Muk is transcribed, it yields a long (>2 kb) dsRNA hairpin structure that is processed to yield siRNAs. The connection between TEs and siRNAs has led to the proposal that origination from TEs distinguishes siRNAs from miRNAs (BARTEL 2004). However, as reported here and elsewhere (SMALHEISER and TORVIK 2005), more and more TE-derived miRNAs are being discovered.

The model of miRNA emergence from MITEs that we propose here (Figure 5.2) suggests a way that miRNAs could have evolved from TE encoded siRNAs. One possible source of the TE encoded dsRNAs that serve as siRNA precursors is snap back panhandle structures between TIRs of autonomous DNA-type elements (VASTENHOUW and PLASTERK 2004). Such panhandle structures would include long internal loop regions that correspond to the internal open reading frames that are lost when autonomous elements

are converted to non-autonomous MITE derivatives. MITEs retain the TIRs, and those same TIRs that were processed from longer RNAs to form siRNA could be similarly processed to form miRNAs. The shorter hairpin structures formed by MITE transcripts could lead to steric constraints that result in the liberation of only one mature miRNA sequence as opposed to the numerous siRNAs that are produced from longer dsRNAs. In this way, short hairpin loop derived miRNAs may have evolved from TE encoded siRNAs. Many of the extant miRNA genes characterized today may have evolved beyond recognition to their progenitor TEs, while others may have originated from other genomic re-structuring mechanisms that juxtapose short inverted repeats (ALLEN *et al.* 2004).

CHAPTER 6

DUAL CODING OF SIRNAS AND MIRNAS BY PLANT TRANSPOSABLE ELEMENTS

ABSTRACT

Short interfering RNA (siRNA) sequences encoded by transposable elements (TEs) are used to silence expression of the elements in order to defend against the harmful effects of transposition. Recently, our group and others demonstrated that TE sequences can also encode miRNAs that are used to regulate cellular (host) genes. We proposed a specific model whereby miRNAs encoded from short non-autonomous DNA-type TEs, known as MITEs, evolved from corresponding ancestral full-length (autonomous) elements that originally encoded siRNAs. This model predicts that evolutionary intermediates may exist as TEs that encode both siRNAs and miRNAs. We analyzed *Arabidopsis thaliana* and *Oryza sativa* (rice) genomic sequence and expression data to test this prediction. We found that there are in fact a number of examples of individual plant TE insertions that encode both siRNAs and miRNAs. We also show evidence that these dual coding TEs can be expressed as read-through transcripts from the intronic regions of spliced RNA messages. These TE-transcripts can fold to form the hairpin (stem-loop) structures characteristic of miRNA genes along with longer double stranded RNA regions that are typically processed as siRNAs. Taken together with a recent study showing Drosha independent processing of miRNAs from Drosophila introns, our results indicate that ancestral miRNAs could have evolved from TEs prior to the full elaboration of the miRNA biogenesis pathway. Later, as the specific miRNA

biogenesis pathway evolved, and numerous other expressed inverted repeat regions came to be recognized by the miRNA processing endonucleases, the host gene related regulatory functions of miRNAs emerged. In this way, host genomes were afforded an additional level of regulatory complexity as a by-product of TE defense mechanisms. The siRNA-to-miRNA evolutionary transition is representative of a number of other regulatory mechanisms that evolved to silence TEs and were later co-opted to serve as regulators of host gene expression.

INTRODUCTION

The phenomenon of RNA-mediated gene regulation was originally discovered in plants (MATZKE and MATZKE 2004). Plant biologists found that posttranscriptional gene silencing (PTGS) seemed to involve RNA or DNA sequence interactions between transgenes, or transgenes and homologous plant genes, which led to sequence-specific RNA degradation (DE CARVALHO *et al.* 1992; NAPOLI *et al.* 1990; VAN BLOKLAND *et al.* 1994; VAN DER KROL *et al.* 1990). It soon became apparent that plant RNA viruses could also stimulate PTGS. Transgenic tobacco plants that expressed a truncated form of a viral coat gene recovered from initial infection with the virus and ultimately became resistant (LINDBO *et al.* 1993). This resistance was found to be conferred through degradation of viral RNA. Subsequently, PTGS was shown to serve as a natural mechanism employed by plants to defend against viral infection (COVEY *et al.* 1997; RATCLIFF *et al.* 1997). Ultimately, these findings led to the notion that a number of plant gene silencing mechanisms initially evolved as defense mechanisms against invading genetic elements (MATZKE *et al.* 2000).

The broader significance of RNA-mediated gene regulation became widely apparent only later, when the specific role of double stranded RNA (dsRNA) in RNA interference (RNAi) was elucidated for *Caenorhabditis elegans* (FIRE *et al.* 1998). RNAi in *C. elegans* was related to genome defense mechanisms by studies showing that RNAi deficient mutants lost the ability to silence Tc*1* transposable elements (TEs) in the germline (KETTING *et al.* 1999; TABARA *et al.* 1999). The mechanism behind RNAi-based silencing of *C. elegans* TEs was found to be based on the production of dsRNAs from the terminal inverted repeat (TIR) sequences found at the ends of Tc*1* elements (SIJEN and PLASTERK 2003). This work demonstrated that RNAi is initiated by read-through transcription of full-length Tc*1* elements, which then fold into 'snap-back' structures with the complementary sequences of the TIRs bound as dsRNA (Figure 6.1). These dsRNA TIR sequences are processed by the RNAi enzymatic machinery to yield short interfering RNAs (siRNAs) that silence expression via mRNA degradation of the transposase gene required for Tc*1* transposition. The sequence-specificity of the mRNA degradation is caused by binding of the TIR-derived single stranded siRNAs to complementary sequences of the transposase encoding mRNA. Later, TE-encoded siRNAs were shown to silence the highly active *MuDR* TE family in maize (SLOTKIN *et al.* 2005). In light of the ability to defend against viral infection and TE mobilization, RNAi has been considered as an immune system for the genome (PLASTERK 2002).

As described above, the connection between the siRNA molecules that mediate RNA-based gene silencing and TEs, or viruses, has been appreciated since PTGS and RNAi were first studied. MicroRNAs (miRNAs) are a related class of short RNA molecules with an analogous functional role in RNAi (AMBROS 2004; BARTEL 2004).

miRNAs are processed from the dsRNA regions of short, ~70-90bp, stem-loop (hairpin) RNA structures by the same endonuclease, Dicer (or Dicer-like in plants), which cleaves siRNAs from longer dsRNA sequences. A connection between TEs and miRNAs was more recently established when a number of miRNA genes were found to be derived from TE sequences (BORCHERT *et al.* 2006; METTE *et al.* 2002; PIRIYAPONGSA and JORDAN 2007; PIRIYAPONGSA *et al.* 2007a; SMALHEISER and TORVIK 2005).

In the human genome, a group of related miRNA genes was found to be derived from the Made1 family of TEs (PIRIYAPONGSA and JORDAN 2007). Made1 elements (MORGAN 1995; OOSUMI *et al.* 1995; SMIT and RIGGS 1996) are members of a specific class on DNA-type TEs known as miniature inverted-repeat transposable elements (MITEs) (BUREAU and WESSLER 1992; BUREAU and WESSLER 1994). MITEs are short non-autonomous derivatives of full-length DNA-type elements (FESCHOTTE and MOUCHES 2000; FESCHOTTE *et al.* 2002b). Full-length DNA-type elements are typically several kb in length and contain a single open reading frame, which encodes the transposase enzyme that catalyzes transposition, flanked by two TIR sequences on either end of the elements (Figure 6.1A). As is the case with the Tc*1* elements of *C. elegans*, full-length transcripts of DNA-type elements can fold into 'snap-back' structures with the two TIRs forming a dsRNA region (Figure 6.1B). This dsRNA region can be processed to yield siRNAs that silence expression of the elements. MITEs are shorter sequences of ~80-500bp, which lack the internal ORF of full-length elements but retain the TIRs (Figure 6.1C). So MITEs are closer to being palindromes, and read through transcription of MITEs will lead to RNA sequences that can fold into hairpin structures reminiscent of

the pre-miRNA the sequences processed by Dicer to yield mature miRNAs (Figure 6.1D).

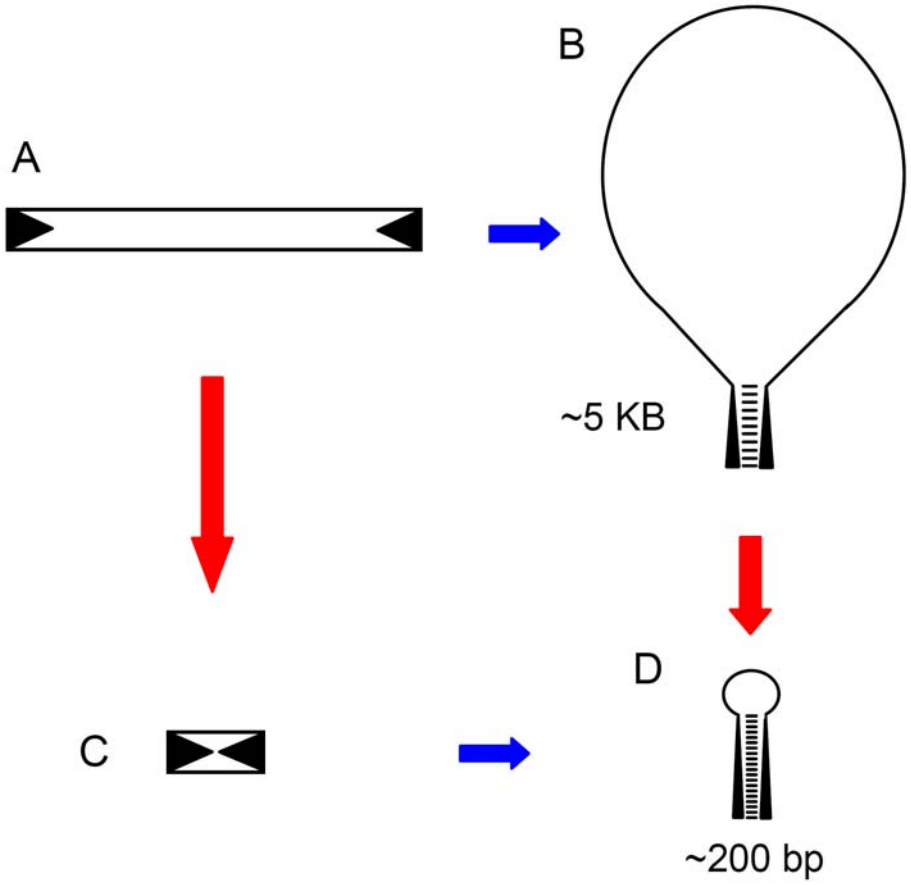

Figure 6.1: Model for the TE-based siRNA-miRNA evolutionary transition. (A) Full-length DNA-type element with terminal inverted repeats (TIRs) flanking a long open reading frame. (B) Snap-back secondary structure of the full-length element with TIRs bound as dsRNA. (C) MITE, a non-autonomous derivative of a full-length DNA-type element, containing TIRs and a small internal region. (D) Hairpin (stem-loop) secondary structure formed by a MITE RNA.

The relationship between full-length DNA-type elements and siRNAs on the one hand, and MITEs and miRNAs on the other, led us to propose a specific model for how miRNAs could have evolved from siRNA encoding TEs in a step-wise manner

(PIRIYAPONGSA and JORDAN 2007). As illustrated in Figure 6.1, our model posits that siRNAs were first processed from the two TIRs of full-length elements bound as dsRNA. Later, as derivative MITEs evolved from full-length elements and proliferated in the genome, the same RNA endonucleolytic processing machinery cleaved the dsRNA from the hairpin stem regions yielding mature miRNA sequences. A corollary prediction of our model holds that evolutionary intermediates may exist as TE sequences that encode both siRNAs and miRNAs. We tested the prediction of dual coding siRNA-miRNA TEs using a computational analysis of genome sequences, annotation and expression data from the plants *Arabidopsis thaliana* and *Oryza sativa* (rice).

RESULTS

We searched Arabidopsis and rice genome sequence and expression data (see Methods) to determine whether there are individual TE insertions that encode both siRNA and miRNA sequences. The Arabidopsis and rice genome sequences, along with their functional genomic datasets, afford several specific advantages for this kind of search. Both model species have been studied extensively, particularly by biologists interested in TEs, and accordingly their TEs are relatively well characterized. In addition, RNA expression levels for Arabidopsis and rice genes have been extensively characterized using the high-throughput massively parallel signature sequencing (MPSS) technique (BRENNER *et al.* 2000a; BRENNER *et al.* 2000b). The original MPSS technique was later modified to characterize small RNA sequences such as siRNAs and miRNAs (LU *et al.* 2006). MPSS for short RNAs yields many thousands of sequence tags that can be unambiguously mapped to the Arabidopsis or rice genomes to determine where mature siRNAs and miRNAs are encoded.

The miRBase Sequence Database (GRIFFITHS-JONES *et al.* 2006) contains genome annotations for experimentally characterized miRNA gene sequences from a number of species including Arabidopsis and rice. Release 10.0 of miRBase contains 184 Arabidopsis and 231 rice miRNAs. We compared the genomic locations of these miRNAs to the locations of TEs annotated using the RepeatMasker program. 12 Arabidopsis miRNAs (6.5%) and 83 rice miRNAs (35.9%) were found to be co-located with TE sequences (Table E.1). 10 out of 12 TE co-located Arabidopsis miRNA sequences and 38 out of 83 TE co-located rice miRNA sequences share 100% of their sequences with TEs. The TE sequences were all annotated based on RepeatMasker scores well above the threshold for false positives (average SW score=20,357). In other words, these data represent unequivocal cases of plant miRNA genes that have been derived from TE sequences (Table 6.1). These miRNAs are derived from members of a variety of TE sequence families including gypsy- and copia-like LTR retroelements, but the vast majority are encoded by the short non-autonomous DNA-type transposable elements known as MITEs. MITE-derived miRNAs are particularly enriched in rice consistent with the genomic abundance of MITEs in this species (JIANG *et al.* 2004).

Table 6.1: Plant miRNA genes derived from TEs

Name[a]	Accn[b]	Coords[c]	TE[d]	TE size[e]
ath-MIR855	MI0005411	chr2:4681509-4681780(+)	Athila4B_LTR (LTR/Gypsy)	fragment
ath-MIR416	MI0001427	chr2:7015602-7015681(+)	Vandal1 (DNA/MuDR)	fragment
ath-MIR405a	MI0001074	chr2:9642037-9642193(-)	SIMPLEHAT2 (DNA/hAT)	fragment
ath-MIR405d	MI0001077	chr4:2789653-2789738(-)	SIMPLEHAT2 (DNA/hAT)	fragment
ath-MIR401	MI0001070	chr4:5020234-5020483(-)	Athila4B_LTR (LTR/Gypsy)	fragment

Table 6.1 continued

Name[a]	Accn[b]	Coords[c]	TE[d]	TE size[e]
ath-MIR854b	MI0005413	chr5:11341600-11341820(-)	Athila6A_I (LTR/Gypsy)	intact
ath-MIR854d	MI0005415	chr5:11707091-11707311(-)	Athila6A_I (LTR/Gypsy)	intact
ath-MIR854c	MI0005414	chr5:11855326-11855546(+)	Athila6A_I (LTR/Gypsy)	intact
ath-MIR854a	MI0005412	chr5:11864949-11865169(+)	Athila6A_I (LTR/Gypsy)	intact
ath-MIR405b	MI0001075	chr5:20649740-20649863 (+)	SIMPLEHAT2 (DNA/hAT)	fragment
osa-MIR439a	MI0001691	chr1:20206990-20207082(+)	MuDR4_OS (DNA/MuDR)	fragment
osa-MIR814a	MI0005239	chr1:22701877-22701973(+)	STOWAWAY47_OS (DNA/Stowaway)	intact
osa-MIR812a	MI0005233	chr1:34273999-34274232(+)	STOWAWAY51_OS (DNA/Stowaway)	intact
osa-MIR819a	MI0005252	chr1:41534243-41534367(+)	STOWAWAY1_OS (DNA/Stowaway)	intact
osa-MIR812b	MI0005234	chr2:1936324-1936493(-)	STOWAWAY51_OS (DNA/Stowaway)	intact
osa-MIR818b	MI0005248	chr2:4007187-4007299(+)	STOWAWAY15-2_OS (DNA/Stowaway)	intact
osa-MIR806b	MI0005211	chr2:5044109-5044323(-)	TREP215 (DNA/Stowaway)	intact
osa-MIR814c	MI0005241	chr2:10889670-10889752(-)	STOWAWAY47_OS (DNA/Stowaway)	fragment
osa-MIR817	MI0005246	chr2:12276361-12276443(-)	ENSPM3_OS (DNA/En-Spm)	fragment
osa-MIR807b	MI0005218	chr2:24481931-24482076(-)	ECR (DNA/Tourist)	intact
osa-MIR814b	MI0005240	chr2:26335342-26335415(+)	STOWAWAY47_OS (DNA/Stowaway)	intact
osa-MIR819d	MI0005255	chr3:10848548-10848699(-)	STOWAWAY1_OS (DNA/Stowaway), STOWAWAY10_OS (DNA/Stowaway)	intact
osa-MIR821a	MI0005266	chr3:22928833-22929106(+)	ENSPM3_OS (DNA/En-Spm), OSTE22 (DNA)	fragment
osa-MIR443	MI0001708	chr3:29972009-29972156(+)	STOWAWAY47_OS (DNA/Stowaway)	intact
osa-MIR420	MI0001440	chr4:6098543-6098697(+)	TRUNCATOR2_OS (LTR/Gypsy)	intact
osa-MIR416	MI0001436	chr4:17268776-17268884(+)	CPSC3_LTR (LTR/Copia)	intact
osa-MIR807c	MI0005219	chr4:23886344-23886527(+)	ECR (DNA/Tourist)	intact
osa-MIR442	MI0001707	chr4:32149607-32149839(+)	OLO24B (DNA/Tourist)	almost full-length

Table 6.1 continued

Name[a]	Accn[b]	Coords[c]	TE[d]	TE size[e]
osa-MIR819f	MI0005257	chr4:35070636-35070779(-)	STOWAWAY50_OS (DNA/Stowaway)	intact
osa-MIR819g	MI0005258	chr5:28003948-28004094(+)	STOWAWAY1_OS (DNA/Stowaway)	intact
osa-MIR819h	MI0005259	chr6:10052973-10053127(-)	STOWAWAY50_OS (DNA/Stowaway), SZ-66LTR (LTR/Gypsy)	intact
osa-MIR811a	MI0005230	chr6:13901553-13901742(+)	TAMI2 (DNA)	intact
osa-MIR812c	MI0005235	chr6:26259310-26259473(+)	STOWAWAY9_OS (DNA/Stowaway)	intact
osa-MIR821b	MI0005267	chr7:16415531-16415817(+)	OSTE22 (DNA), TNR3_OS (DNA/En-Spm)	fragment
osa-MIR812d	MI0005236	chr7:22393529-22393681(+)	STOWAWAY44_OS (DNA/Stowaway)	intact
osa-MIR445a	MI0001709	chr7:28117531-28117798(+)	NDNA2TNA_OS (DNA/Tourist)	intact
osa-MIR818e	MI0005251	chr7:28152738-28152962(-)	STOWAWAY21_OS (DNA/Stowaway)	intact
osa-MIR531	MI0003204	chr8:1214013-1214093(-)	SC-1_int-int (LTR/Copia)	fragment
osa-MIR812e	MI0005237	chr8:16268303-16268472(+)	STOWAWAY44_OS (DNA/Stowaway)	intact
osa-MIR821c	MI0005268	chr8:19792287-19792552(-)	ENSPM3_OS (DNA/En-Spm), OSTE22 (DNA)	almost full-length
osa-MIR811b	MI0005231	chr10:2372014-2372203(+)	TAMI2 (DNA)	intact
osa-MIR439b	MI0001692	chr10:5338996-5339055(+)	MuDR4_OS (DNA/MuDR)	fragment
osa-MIR816	MI0005245	chr10:21478646-21478722(+)	STOWAWAY47_OS (DNA/Stowaway)	almost full-length
osa-MIR806g	MI0005216	chr10:22588399-22588638(+)	TREP215 (DNA/Stowaway)	intact
osa-MIR811c	MI0005232	chr11:5200383-5200541(-)	TAMI2 (DNA)	fragment
osa-MIR813	MI0005238	chr11:23113437-23113639(+)	NDNA1TNA_OS (DNA/Tourist)	fragment
osa-MIR531	MI0003204	chr11:26423868-26423948(+)	SC-1_int-int (LTR/Copia)	intact
osa-MIR809h	MI0005228	chr12:5776955-5777088(+)	STOWAWAY1_OS (DNA/Stowaway)	intact

[a]miRBase database miRNA names
[b]miRBase database miRNA accessions
[c]genomic location coordinates of co-located TE sequences
[d]name, class and family of the TE sequences

[e]size of TE sequences which encode miRNA genes: intact (full-length element), almost full-length (≥80% of full-length element consensus sequence), fragment (<80% of full-length element consensus sequence).

miRBase was used to count the number of orthologs for each Arabidopsis and rice miRNA. TE-derived miRNA genes in Arabidopsis and rice have fewer orthologs on average (0.07), *i.e.* they are less evolutionarily conserved, than non repetitive miRNAs (3.0) and the difference is highly significant (Student's t-test; t=18.8 df=413 P=2e-57). This is similar to what is seen for many mammalian TE-derived miRNAs (PIRIYAPONGSA *et al.* 2007a) and is consistent with the fact that TEs represent the most lineage-specific and rapidly evolving sequences in eukaryotic genomes (MARINO-RAMIREZ *et al.* 2005). On the one hand, this may suggest that caution is warranted when evaluating TE-derived plant miRNAs since they are not conserved (AMBROS *et al.* 2003). However, there are a number of *bona fide* miRNAs that are not evolutionarily conserved (BENTWICH *et al.* 2005). The low conservation of TE-derived miRNAs can be taken to imply that the regulatory effects exerted by TE-derived miRNAs may be relevant for species-specific differences in gene expression (PIRIYAPONGSA *et al.* 2007a).

In addition to using miRBase to characterize TE-derived miRNAs, we searched the literature to confirm TE-derived plant miRNA genes with documented effects on the expression of host genes. There are five TE-derived miRNAs uncovered here (ath-MIR854a-d & ath-MIR855 in Table 6.1), including a repetitive family derived from dispersed LTR sequences, with experimentally characterized effects on the regulation of Arabidopsis genes (ARTEAGA-VAZQUEZ *et al.* 2006). First of all, mature ath-MIR854 sequences were found to be absent in plants with mutant alleles for three genes critical to

miRNA biogenesis: *Dicer-like1* (*dcl1*), *Hyponastic leaves1* (*hyl1*) and *HUA Enhancer1* (*hen1*). However, ath-MIR854 expression was found in mutants of the *RNA-dependent RNA polymerase2* gene, which is required for siRNA processing. Together, these results indicate that ath-MIR854 is processed specifically as an miRNA. The mature sequences of ath-MIR854 and ath-MIR855 have multiple binding sites in the 3' untranslated region (UTR) of the Oligouridylate binding protein1b gene (*UBP1b*), which encodes a heterogenous nuclear RNA binding protein. The *UBP1b* 3'UTR mRNA: miRNA interactions resemble those that lead to translational repression and/or mRNA cleavage in mammals. The ability of these TE-derived miRNAs to repress expression of *UBP1b* was demonstrated by using the 3' UTR of the gene in a reporter protein expression assay. Mature ath-MIR854 and ath-MIR855 sequences are expressed in rosette leaves and flowers but absent in cauline leaves. Accordingly, the 3' UTR of *UBP1b* can repress protein reporter expression in rosette leaves and flowers not in cauline leaves. Furthermore, comparison of mRNA versus protein expression for the reporter indicated that the ath-MIR854a-d and ath-MIR855 genes exert their effects at the translational level.

We used expression data taken from the Arabidopsis MPSS Plus (MEYERS *et al.* 2004) and Rice MPSS Plus (NOBUTA *et al.* 2007) databases to evaluate whether any of the TEs that encode miRNA genes are also processed to yield siRNA sequences. The siRNA MPSS sequence tags were unambiguously mapped, using 100% tag-TE sequence identity, to the TEs that were found to encode miRNAs. 8 of the Arabidopsis TEs that encode miRNAs and 13 of the miRNA encoding rice TEs were found to encode siRNA sequences as well (Table 6.1).

We also explored the possibility that TE-derived miRNAs and siRNAs are passively transcribed as part of longer host gene RNA messages. To do this, genome-wide EST and mRNA maps for Arabidopsis and rice were examined to look for cases where dual coding miRNA-siRNA TE sequences are located within the start and stop coordinates of spliced transcripts. We were able to find several examples of such cases, 1 for Arabidopsis and 2 for rice (Figure 6.2).

Finally, all of the miRNA-siRNA dual coding TE sequences were folded to predict their secondary structures (Figure E.1). The predicted secondary structures show long double stranded regions that correspond to the locations of mapped siRNA sequence tags along with stem-loop regions characteristic of known miRNA gene structures. The MITE encoded secondary structures are particularly striking in the sense that they form long, almost perfect, hairpins possessing extensive double stranded regions (Figure 6.3 and Figure E.1). These folding patterns are based on the sequence complementarity between the terminal inverted repeats (TIRs) encoded by this class of TEs.

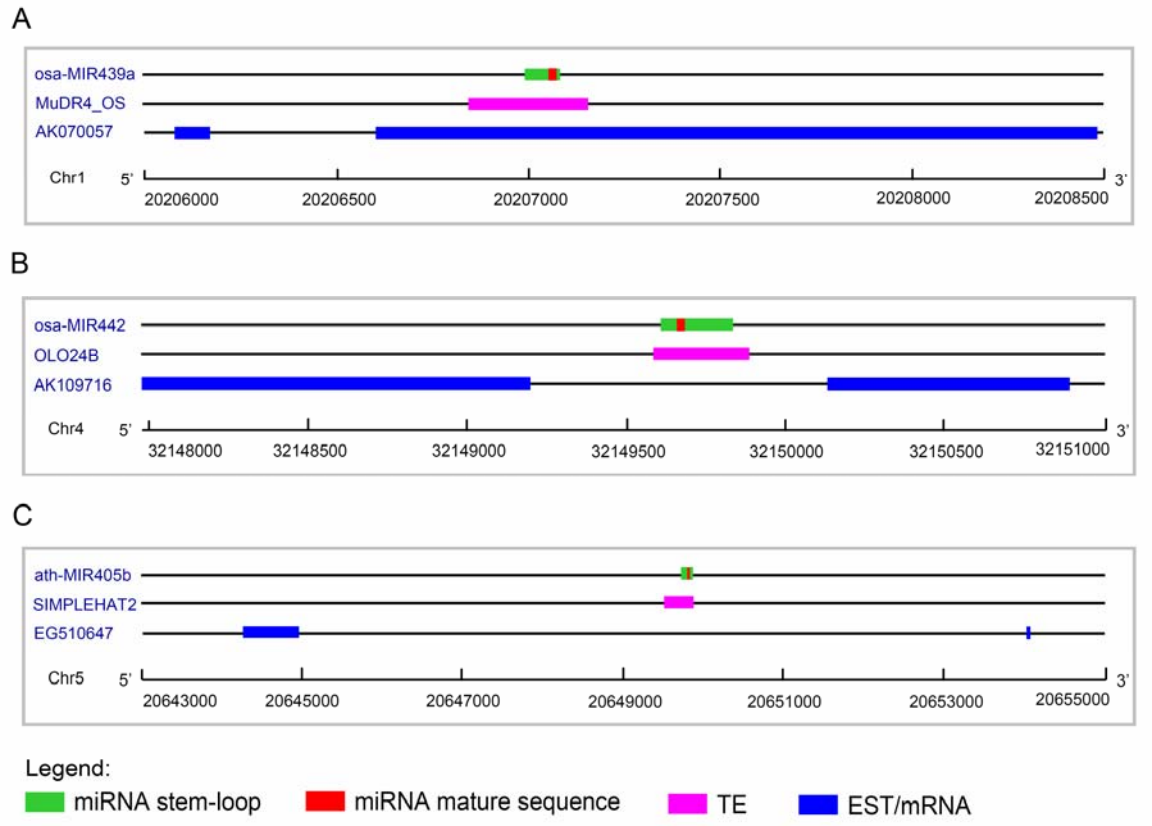

Figure 6.2: Genomic structure and expression of TE-derived miRNAs. The schematic diagrams representing the co-location between miRNA genes, TEs and EST/mRNAs (see legend) are shown for (A) osa-MIR439a (B) osa-MIR422 and (C) ath-MIR405b.

DISCUSSION

Our analysis of Arabidopsis and rice genomic data revealed the existence of TE sequences that encode both siRNAs and miRNAs. We believe that the dual coding capacity for small regulatory RNAs by plant TEs reflects an evolutionary connection between related mechanisms of RNAi. This notion is based on the recent discovery of a family of human miRNA genes derived from MITEs, which led us to propose a step-wise model for the evolution of miRNAs from TEs that originally encoded siRNAs (PIRIYAPONGSA and JORDAN 2007). The expression of siRNAs from autonomous DNA-type elements is known to be based on read-through transcription of full-length elements

(SIJEN and PLASTERK 2003), and our model is based on read-through transcription of shorter non-autonomous MITEs (Figure 6.1). MITEs retain the TIR sequences of autonomous DNA-type elements but do not encode any open reading frame between the TIRs. As such, MITEs are made up mostly of TIR sequences, *i.e.* they are palindromic, and when expressed as read-through transcripts, they will fold to form hairpin structures similar to those of miRNA genes (BARTEL 2004). Apparently, these MITE-derived hairpins can be processed to yield functionally relevant mature miRNA sequences (PIRIYAPONGSA and JORDAN 2007; PIRIYAPONGSA *et al.* 2007a). Consistent with this model, our results demonstrate that several of the siRNA-miRNA encoding TEs found in plants are in fact expressed as read-through transcripts by virtue of their presence in the introns of spliced RNA messages (mRNAs/ESTs in Figure 6.2). After TE-containing introns are spliced from the mRNAs, they can fold to form the kinds of structures recognized by the endonucleases involved in RNAi (Figure 6.3 and Figure E.1).

In addition to their palindromic sequence-structure characteristics, MITEs are also distinguished by their preference for insertion into gene rich regions (MAO *et al.* 2000; ZHANG *et al.* 2000). Taken together with their genomic abundance (JIANG *et al.* 2004), this means that thousands of MITEs will be expressed as read-through transcripts as required by our model. The particular enrichment of MITEs in plant gene regions has been taken to suggest that they play some functional role for their host genomes. Our results, and our model of miRNA evolution via the autonomous TE-to-MITE transition, suggest that the host relevant function of MITEs is related, at least in part, to RNA-mediated gene regulation.

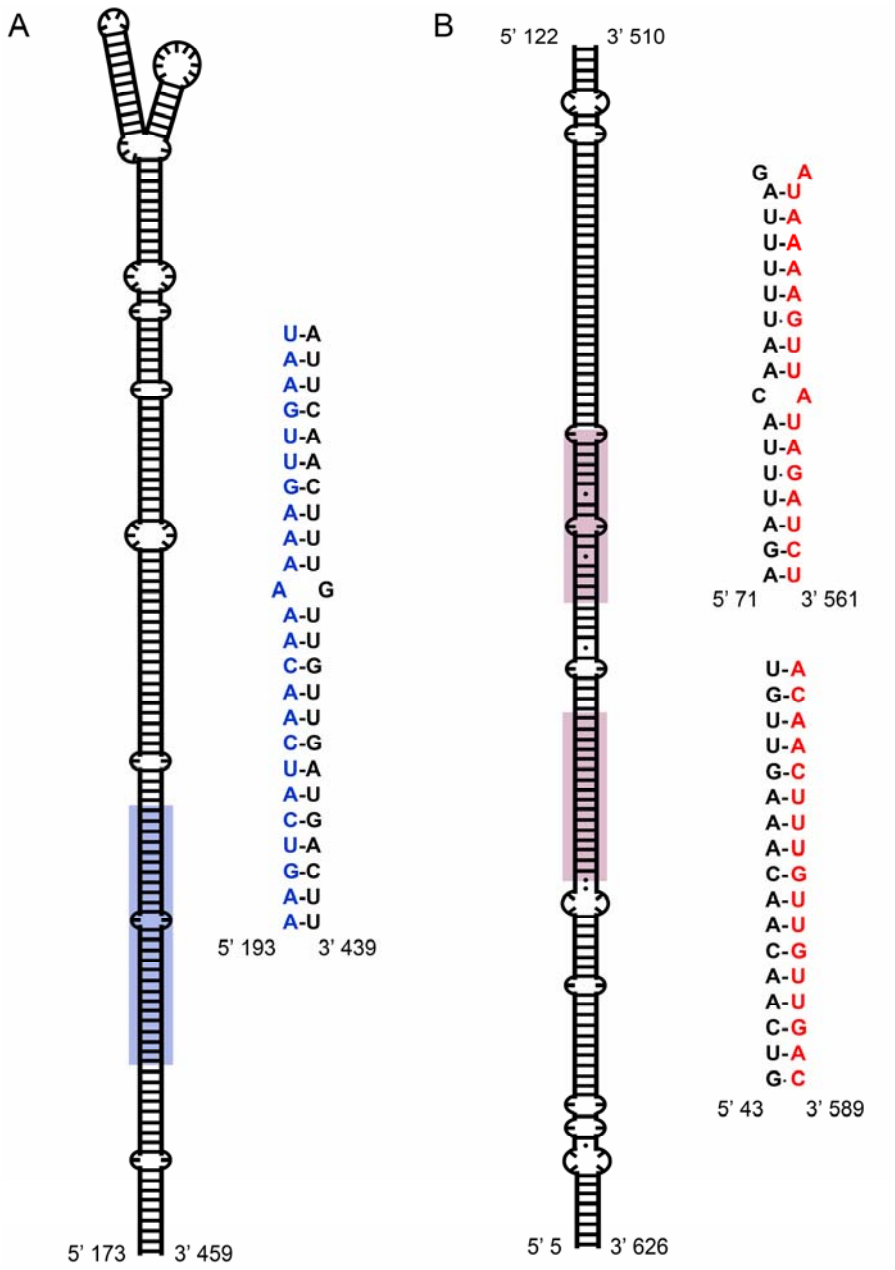

Figure 6.3: RNA secondary structure and sequences of an siRNA-miRNA dual encoding MITE sequence. Partial predicted secondary structures of a read-through transcript of a MITE encoding both siRNA and miRNA sequences are shown for the rice miRNA gene osa-MIR821b. (A) A schematic of the miRNA stem-loop region along with the miRNA mature sequence. The location of the miRNA mature sequence in the stem-loop is indicated with blue shading, and the mature miRNA sequence residues are shown in blue. (B) A schematic of the double-stranded RNA region that is cleaved to yield siRNAs. The locations of the siRNA signatures in the sequence are indicated with red shading, and the siRNA signature sequences are shown in red. Note that the entire secondary structure for this MITE is shown in Figure E.1R.

There is recent evidence from Drosophila in support of the notion that miRNAs can be processed from the introns of expressed genes in a manner similar to that which we propose for TE-derived miRNA genes (RUBY *et al.* 2007). Spliced introns that can fold into pre-miRNA like structures and be cleaved to yield mature miRNA sequences are called 'mirtrons'. Interestingly, the processing of mirtrons to yield mature miRNAs does not rely on the RNA endonuclease Drosha. Drosha, or its plant functional analog Dicer-like1, is the enzyme that cleaves the longer pri-miRNA sequence near the base of the stem region in the nucleus to yield the pre-miRNA hairpin, which is then cleaved by Dicer to liberate the mature miRNA. Similar to the processing of mirtrons, PTGS via siRNAs does not require Drosha since dsRNAs are processed by Dicer alone to yield siRNAs (BARTEL 2004). Thus, the work of Ruby et al. (2007) indicates that miRNAs can arise in any organism that possesses both spliceosomal introns and PTGS via siRNAs; this was taken to suggest that miRNAs may have emerged in ancient eukaryotes prior to the evolution of the complete miRNA biogenesis pathway. Our model points to MITEs as a potential source for the evolution of such ancient miRNAs, processed by Dicer (or Dicer-like1) alone, from full-length TEs that previously encoded siRNAs only. Consistent with an ancient origin of miRNAs from TEs, full-length DNA-type elements and MITEs are widely distributed among eukaryotes (FESCHOTTE *et al.* 2002b), indicating that they were likely to be present in ancestral eukaryotic species. In addition, a recent phylogenetic analysis of the miRNA biogenesis enzymes indicates that Dicer is the more ancient of the endonucleases involved the processing of mature miRNAs, with Drosha having evolved more recently along the animal evolutionary lineage (CERUTTI and CASAS-MOLLANO 2006).

dsRNA sequences are processed to yield multiple siRNAs from a given stretch of sequence, while pre-miRNA hairpins are cleaved into a single distinct mature miRNA sequence (BARTEL 2004). This may be related to the steric hindrance entailed by the substantially shorter hairpin structures that are processed to yield miRNAs. Over evolutionary time, once the endonucleolytic machinery became tuned to the structural characteristics, and limited spacing, of the MITE-encoded hairpins, then it would have been able to recognize any number of hairpin structures that are formed when genomic sequences with inverted repeats are expressed as read-through transcripts. Indeed, this has been shown to be important in Arabidopsis where miRNA genes evolved via local inverted duplication events, which generated sequences capable of folding back into hairpin structures when expressed (ALLEN *et al.* 2004). In this way, MITEs could have stimulated the RNAi biogenesis enzymes to process non TE-related hairpin structures to yield miRNAs with host gene regulatory functions.

Relatively ancient siRNA sequences originally evolved as defense mechanisms against genomic invaders, such as viruses and TEs, and genome defense appears to remain the primary function of this class of regulatory sequence. On the other hand, miRNAs are evolutionarily emergent regulators, and accordingly they function primarily to regulate host genes. The siRNA to miRNA evolutionary transition is one of a growing number of examples (LIPPMAN *et al.* 2004; MATZKE *et al.* 2000; MCDONALD *et al.* 2005; YODER *et al.* 1997) of gene silencing mechanisms that were originally employed to defend against TE proliferation and were later co-opted to serve the regulatory needs of the host organism (PIRIYAPONGSA *et al.* 2007a).

MATERIALS AND METHODS

The *Arabidopsis thaliana* genomic sequence was obtained from the National Center for Biotechnology (NCBI) genome assembly/annotation projects ftp site (ftp://ftp.ncbi.nih.gov/genomes/Arabidopsis_thaliana). The *Oryza sativa* (rice) genomic sequence was taken from release 4.0 of The Institute for Genomic Research (TIGR) rice genome annotation database (OUYANG *et al.* 2007) (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/). The genome locations of different classes of TEs in Arabidopsis and rice genomes were identified by using the RepeatMasker program (SMIT *et al.* 1996-2004) to compare genomic sequences against the species-specific Repbase libraries (JURKA 2000; JURKA *et al.* 2005) of TE consensus sequences. The Smith-Waterman (SW) algorithm (SMITH and WATERMAN 1981) was used with RepeatMasker to do local pairwise comparisons of genome-against-TE consensus sequences and to score the resulting alignments. The genome locations and identities of experimentally characterized Arabidopsis and rice miRNA gene sequences were taken from release 10.0 of the miRBase database (GRIFFITHS-JONES *et al.* 2006) (http://microrna.sanger.ac.uk/sequences/).

Arabidopsis and rice MPSS small RNA signatures were downloaded from the Arabidopsis MPSS Plus database (MEYERS *et al.* 2004) (http://mpss.udel.edu/at/) and the Rice MPSS Plus database (NOBUTA *et al.* 2007) (http://mpss.udel.edu/rice/). The signatures matching to tRNAs, rRNAs, snRNAs or snoRNAs were not included in the data set we used. Only the small RNA signatures of size 17 to 25 bp were chosen for the analysis. For each species, the small RNA signatures were divided into two groups:

miRNA signatures and siRNA signatures according the organism-specific MPSS database annotations.

The genome locations of TEs and miRNAs were compared to identify the co-located TEs and miRNA gene sequences in both species. Post-processing of RepeatMasker annotations were done such that the continuous TE sequences of the same family, which are oriented in the same direction on the genome, were counted as the same TE sequence. TE sequences which encoded entire miRNA gene sequences were searched for the presence of small RNA signatures using the vmatch program (ABOUELHODA *et al.* 2004) demanding 100% sequence identity between the TE sequences and siRNA tags. The TE sequences that completely covered miRNA gene sequences and contained siRNA signatures outside the miRNA gene regions were chosen for further analysis. These TE sequences were folded using the program RNAfold from the Vienna RNA package (HOFACKER *et al.* 1994) and their secondary structures were visualized by xrna program (http://rna.ucsc.edu/rnacenter/xrna/xrna.html). The potential of TE-derived miRNAs and siRNAs to be processed from read-through transcripts was assessed via the analysis of EST and mRNA data. EST and mRNA sequences mapped to the Arabidopsis genome sequence were obtained from NCBI genome assembly/ annotation projects (ftp://ftp.ncbi.nih.gov/genomes/Arabidopsis_thaliana/GNOMON). Mapped rice EST, full-length cDNA sequences and transcript assemblies were obtained from TIGR rice genome annotation database (OUYANG *et al.* 2007).

CHAPTER 7

CONCLUSION

In summary, this dissertation is composed of five different studies that provide new insights into the same field of biological investigation, namely the contribution of TEs to host gene sequences. The first two studies (CHAPTER 2 and CHAPTER 3) present new results with respect to the extent, evolution, and coding property of TE-derived CDS in the human genome as well as the ability of the different search methods used to detect such sequences. These results contribute significantly to the understanding of molecular domestication events of TEs in the human genome. The detailed analysis of the relationship between TEs and a recently discovered class of non-coding regulatory gene, miRNA, presented in CHAPTER 4 and CHAPTER 5 provides new evidences supporting the hypothesis of TE contributions to regulatory gene evolution. Furthermore, the last study (CHAPTER 6) proposes a novel concept for the possible role of one particular class of TEs, MITEs, as an evolutionary link between miRNAs and closely related siRNAs.

Dismissed for some time as "junk", or "selfish" DNA (DOOLITTLE and SAPIENZA 1980; HICKEY 1982; OHNO 1972; ORGEL and CRICK 1980), TEs are now generally considered as significant contributors to gene and genome evolution (BROSIUS 1999; BROSIUS and GOULD 1992; KAZAZIAN 2004; KIDWELL and LISCH 2001; MAKALOWSKI 2003; MCDONALD 1993; MCDONALD 1999).

In CHAPTER 2, a detailed analysis of exonization events in the human genome associated with one specific class of TE, LTR elements, is reported. 5.8% of human

genes were found to contain sequences derived from LTR retrotransposons. The distribution of these elements in genes shows the preference towards the fixation in gene untranslated regions, which supports the existing concept a major role of LTR elements as a natural source of regulatory sequences. On the other hand, the recruitment of LTR retrotransposon sequence as host CDS is not a frequent event. Several coincidences are necessary to allow for an LTR exonization event that yields a CDS. Only 50 protein coding exons were completely derived from LTR retrotransposon sequences. Finally, as shown in the part of this study, the evolutionary analysis using new experimental evidence elucidates the mechanism of incorporation of LTR sequence into an alternatively spliced exon of *IL22RA2* gene and estimates the emergence time of this exon in great ape species. A single mutation in the proto-splice site was hypothesized to cause the recruitment of this novel exon prior to the divergence of orangutans and humans from a common ancestor.

Although a number of large-scale analyses have been used to identify the instances of TE-derived CDSs in human genome (BRITTEN 2006; GOTEA and MAKALOWSKI 2006; LANDER *et al.* 2001; NEKRUTENKO and LI 2001; PAVLICEK *et al.* 2002), the actual proportion of human CDS that have evolved from TEs remains to be defined. In particular, it is unclear whether non-autonomous TEs that do not encode any protein can indeed provide protein coding sequences after becoming exonized (PAVLICEK *et al.* 2002). The ascertainment biases related to different sequence similarity search methods used and the potential of TEs to contribute protein coding sequences are evaluated in CHAPTER 3. The profile-based search methods (*i.e.* HMM) show a beneficial combination of sensitivity and selectivity compared to other search methods

used. However, possibly due to the superior selectivity of the profile-based methods, not many novel cases were detected when these methods were similarly applied to large-scale datasets of experimentally characterized proteins. In general, the different search methods are found to be complementary, and combined search approaches are needed to systematically check any data set for all potential TE-CDS associations. The codon based analysis of exonized TE sequences implies that many of the sequences derived from non-coding TEs, such as Alu elements, are not likely to actually encode any protein. The apparent low coding potential of Alu-derived exons may also reflect the fact that these sequences have a relatively recent evolutionary origin as exons and thus have not had enough time to establish sequence periodicities that resemble other coding sequences. The lack of protein coding capacity does not directly imply the non-functionality of exonized TE sequences. Alternatively, they may play a role in post-transcriptional gene regulation, *e.g.* serve as natural anti-sense transcripts as was recently shown (CONLEY *et al.* 2008). The repetitive dispersed nature of exonized TE sequences may provide a mechanism by which they can serve as master regulators with influence over the expression of numerous genes throughout the genome.

In addition to the contribution to host gene coding sequences, TEs are well recognized for their influences on host gene regulation (BROSIUS 1999; HAMDI *et al.* 2000; JORDAN *et al.* 2003; KIDWELL and LISCH 1997; KIDWELL and LISCH 2001; TOMILIN 1999; VAN DE LAGEMAAT *et al.* 2003). Considerable evidence now indicates that small noncoding RNAs can play a major role in regulating eukaryotic gene expression (CULLEN 2002; HUTVAGNER and ZAMORE 2002b). Of particular interest are a class of ~22-nt RNAs: siRNAs and miRNAs (AMBROS *et al.* 2003). The connection between TEs

and siRNAs has led to the proposal that origination from TEs distinguishes siRNAs from miRNAs (BARTEL 2004).

In CHAPTER 4, the evolutionary relationship between miRNAs and TEs is revealed. About 12% of experimentally verified human miRNA genes were shown to originate from TEs. The dispersed repetitive nature of TE sequences provides for the emergence of multiple novel miRNA genes as well as numerous homologous target sites throughout the genome. Overall, TE-derived miRNA genes are less conserved than non TE-derived miRNAs. This result is generally consistent with the observation that TEs are the most lineage-specific, recently evolved sequences in the human genome (LANDER *et al.* 2001). However, there are a number of TE-derived miRNAs that are well-conserved. The majority of these conserved miRNAs are related to the ancient L2 and MIR TE families, which have been shown to be anomalously conserved between the human and mouse genomes although no specific functional role was assigned to them (SILVA *et al.* 2003). At least some of these conserved L2 and MIR fragments were shown to provide miRNA sequences with the potential to regulate numerous human genes. Along with this study, 85 putative miRNA genes were predicted from the set of TE sequences encoding conserved RNA secondary structures. The high frequency (>50%) of TE-encoded conserved secondary structures which were associated with recently identified conserved TE famlies suggest that many conserved TE sequences may encode miRNAs or perhaps other noncoding RNAs.

Out of 55 TE-derived miRNAs analyzed, a miRNA gene family, hsa-mir-548, was found to be derived from Made1 elements, which are non-autonomous DNA-type TEs known as MITEs (CHAPTER 5). The palindromic structure of the Made1 elements

together with their insertion into transcriptionally active genomic regions, points to a specific mechanism by which these sequences can be recognized and processed by the miRNA biogenesis pathway. Made1 elements emerged along the primate evolutionary lineage, and thus represent a natural source of lineage-specific miRNAs, which could in turn be responsible for regulatory phenotypes that contribute to evolutionary diversification between species.

The model of miRNA emergence from MITEs presented here, together with the relationship between full-length DNA-type elements and siRNAs (VASTENHOUW and PLASTERK 2004), leads to the original idea that MITEs may represent an evolutionary link between siRNAs and miRNAs. A step-wise model for the evolution of miRNAs from TEs that originally encoded siRNAs is proposed in CHAPTER 6. In this model, miRNAs evolved from the autonomous TE-encoded siRNA to MITE-encoded miRNA. This model is supported by the presence of evolutionary intermediate TE sequences that encode both siRNAs and miRNAs in the Arabidopsis and rice genomes. These results demonstrate that several of the siRNA-miRNA encoding TEs found in introns can be expressed as read-through transcripts. After TE-containing introns are spliced from the mRNAs, they can fold to form structures recognized by the RNAi enzyme machinery. Spliced introns that can fold into pre-miRNA like structures and be cleaved to yield mature miRNA sequences are called 'mirtrons' (RUBY *et al.* 2007). Because the processing of mirtrons to yield mature miRNAs does not depend on the RNA endonuclease Drosha, miRNAs can arise in any organism that possesses both spliceosomal introns and PTGS via siRNAs. This suggests that miRNAs may have emerged in ancient eukaryotes prior to the evolution of the complete miRNA biogenesis

pathway. Consistent with an ancient origin of miRNAs from TEs, full-length DNA-type elements and MITEs are widely distributed among eukaryotes (FESCHOTTE *et al.* 2002b), indicating that they were likely to be present in ancestral eukaryotic species. In addition, a phylogenetic analysis of the miRNA biogenesis enzymes indicates that Dicer is the more ancient of the endonucleases involved the processing of mature miRNAs, with Drosha having evolved more recently along the animal evolutionary lineage (CERUTTI and CASAS-MOLLANO 2006).

In conclusion, the results from miRNA-TE analysis point to a connection between genome defense mechanisms necessitated by TEs and the emergence of global gene regulatory systems. A number of epigenetic gene silencing mechanisms, such as cytosine methylation (YODER *et al.* 1997), genomic imprinting (MCDONALD *et al.* 2005) and heterochromatin (LIPPMAN *et al.* 2004) are thought to have evolved originally as TE defense mechanisms. Subsequently, these TE silencing mechanisms were co-opted as global regulators to control the phenotypic complexity seen among members of the eukaryotes. The evolutionary transition from siRNA to miRNA through MITEs presented in this study shows one of a growing number of of examples of this scenario. RNA interference by siRNAs is considered to have originally evolved to silence TEs (MATZKE *et al.* 2000; SLOTKIN *et al.* 2005; VASTENHOUW and PLASTERK 2004) and the primary function of siRNA sequence appears to be genome defense against genomic invaders while miRNAs are evolutionarily emergent regulators, and accordingly they function primarily to regulate host genes.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A.1: Features of LRTS-derived protein coding exons

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions [7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
internal protein coding exon													
1	NM_005799*	+/-	34/36*	90(30)	1582	368	MSTD	MaLR[1]	InaD-like protein (*INADL*), transcript variant 3	protein binding (protein-protein interaction mediated by PDZ domains)	AJ224748	N/A	* GT/AG
2	NM_198712*	-/-	4/6*	70	400	265	MSTA	MaLR[1]	prostaglandin E receptor 3 (subtype EP3) (*PTGER3*), transcript variant 2	prostaglandin E receptor activity, ligand-dependent nuclear receptor activity, rhodopsin-like receptor activity	D86097	N/A	* donor (all = AT); acceptor (c & h = AG, r = TG)
3	NM_198713*	-/-	3/5*	76	393	265	MSTA	MaLR[1]	prostaglandin E receptor 3 (subtype EP3) (*PTGER3*), transcript variant 3	prostaglandin E receptor activity, ligand-dependent nuclear receptor activity, rhodopsin-like receptor activity	D86098, AY429108	N/A	* donor (all = GT); acceptor (c & h = AG, r = TG)
4	NM_020161*	-/-	2/3	73	150	382	MSTB2	MaLR[1]	hypothetical protein DKFZp547H025 (*DKFZp547H025*)	folic acid binding, reduced folate carrier activity	AK055355, AL359944, CR749679	DA513601	* donor (all = GT); acceptor (c & h = AG, r = GG)
5	NM_007072*	+/-	8/10	51(17)	414	217	LTR33	ERVL[1]	HERV-H LTR-associating 2 (*HHLA2*)	unknown	BC035971, AF126162, AK000692, AK027132	DB232201, BG283385	GT/AG
6	NM_001025468	+/+	2/3	95	97	408	MSTA	MaLR[1]	chromosome 3 open reading frame 47 (*C3orf47*)	undefined	BC093941, AK091470, BC101739	DA499220, BF514892, AL040547, BM709191	* GT/AG

Table A.1 continued

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions [7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
7	NM_145027*	-/+	20/23	51(17)	814	361	MLT1A0	MaLR[1]	kinesin family member 6 (*KIF6*)	ATP binding, nucleotide binding, microtubule motor activity, microtubule-based movement	AL832634, BC022074, AK131471	CA314029, BM696570, BI118088, BG189781, CV024391, BG715118	GT/AG
8	NM_052962*	-/-	4/7*	96(32)	263	445	MSTB2	MaLR[1]	interleukin 22 receptor, alpha 2 (*IL22RA2*), transcript variant 1	interleukin-22 receptor activity, hematopoietin/ interferon-class (D200-domain) cytokine receptor activity	AY040567, AJ313162, AY358737	N/A	* donor (c & h = GT, r = AT); acceptor (all = AG)
9	NM_024728*	+/-	13/15	78(26)	434	519	MLT1E2	MaLR[1]	chromosome 7 open reading frame 10 (*C7orf10*)	transferase activity	AK021870, BC098318	BG678361, CN259576	GT/AG
10	NM_174930	+/+	5/6	89	134	387	MSTA	MaLR[1]	postmeiotic segregation increased 2-like 5 (*PMS2L5*)	ATP binding, damaged DNA binding (in mismatch repair process)	BC027480, D38436, D38501, D38502, D38439	BQ049786, DA486906, BE883788, AA621085, CA442493, AA411808, AA917605, AA282042	* donor (c & h = GT, r = AT); acceptor (all = AG)
11	NM_002679	-/-	7/8	89	297	387	MSTA	MaLR[1]	postmeiotic segregation increased 2-like 2 (*PMS2L2*)	unknown (in mismatch repair process)	AB017005	BM547367, BE703998, BE703901, BE701175, DR001682, AU120696, DR007875, BE163029	* donor (c & h = GT, r = AT); acceptor (all = AG)
12	NM_002679	-/-	2/8	89	297	387	MSTA	MaLR[1]	postmeiotic segregation increased 2-like 2 (*PMS2L2*)	unknown (in mismatch repair process)	AB017005	BM547367, BE703998, BE703901, BE701175, DR001682, DR007875, BE163029, BQ049786	* donor (c & h = GT, r = AT); acceptor (all = AG)

Table A.1 continued

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions [7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
13	NM_002679	-/-	7/8	89	297	387	MSTA	MaLR[1]	postmeiotic segregation increased 2-like 2 (*PMS2L2*)	unknown (in mismatch repair process)	AB017005	BM547367, BE703998, BE703901, BE701175, DR001682, DR007875, BE163029, BQ049786	* donor (c & h = GT, r = AT); acceptor (all = AG)
14	NM_002679	-/-	2/8	89	297	383	MSTA	MaLR[1]	postmeiotic segregation increased 2-like 2 (*PMS2L2*)	unknown (in mismatch repair process)	AB017005, AB017007	BE703998, BE703901, DR001682, AU120696, BG740816, DR007875, BE163029, BQ049786	* donor (c & h = GT, r = AT); acceptor (all = AG)
15	NM_032958*	-/-	3/8*	87(29)	116	384	MSTA	MaLR[1]	DNA directed RNA polymerase II polypeptide J-related gene (*POLR2J2*), transcript variant 2	DNA binding, transferase activity, protein dimerization activity, DNA-directed RNA polymerase activity (a subunit of RNA polymerase II)	BC056864, BC086857, BC071870, CR606303, CR624568, CR596358, CR614811, AJ277740	BM915750, BM559191, BG177266, DA093808, AL535064, DA239331, DA474156, AL556716	* donor (c & h = GT, r = AT); acceptor (all = AG)
16	NM_001023564*	+/-	4/5	94	218	419	MSTC	MaLR[1]	cathepsin L-like protein (*HCTSL-s*)	cysteine-type peptidase activity	AJ851862	N/A	no orthologous exonic sequence in chimpanzee
17	NM_017418*	+/-	6/8	74	70	371	MSTD	MaLR[1]	deleted in esophageal cancer 1 (*DEC1*)	unknown (a candidate tumor suppressor gene for esophageal squamous cell carcinomas located in a region commonly deleted in these carcinomas)	AB022761	N/A	GT/AG

Table A.1 continued

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions[7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
18	NM_001010910*	+/-	2/3	85	102	203	MLT1J	MaLR[1]	hypothetical LOC399706 (*LOC399706*)	iron ion binding, electron transporter activity	AK097673	DB045405	no orthologous exonic sequence in rhesus monkey
19	NM_001548	+/+	2/3*	109	41	402	MLT1B	MaLR[1]	interferon-induced protein with tetratricopeptide repeats 1 (*IFIT1*), transcript variant 2	binding, immume response (interferon-induced protein)	AK092813	DB003897, DA380195, DA671459, DA672179, DA599173, DA676435, DA540285, DA599173	donor (c & h = GT, r = AT); acceptor (all = AG)
20	NM_015430*	-/+	7/12*	51(17)	737	348	MLT1A0	MaLR[1]	regeneration associated muscle protease (*DKFZP586H2123*), transcript variant 1	trypsin activity, peptidase activity, calcium ion binding, chymotrypsin activity	BC038457, BC089434, AK027841	DA475973, BG403264, DA543024, BX439313	GT/AG
21	NM_001001681*	+/+	3/7	104	129	395	MSTD	MaLR[1]	FLJ45300 protein (*FLJ45300*)	undefined	AK127233	DA313959	* GT/AG
22	NM_173580*	+/-	2/3	107	122	133	MER21C	ERV1[1]	chromosome 11 open reading frame 44 (*C11orf44*)	undefined	AK096377	DA748198	no orthologous exonic sequence in rhesus monkey
23	NM_183378	-/+	26/28	107	1134	436	MER34B	ERV1[1]	ovochymase 1 (*OVCH1*)	peptidase activity,serine-type endopeptidase activity	BN000128	N/A	* GT/AG
24	NM_031915*	+/-	5/15	36(12)	719	50	MER34	ERV1[1]	SET domain, bifurcated 2 (*SETDB2*)	DNA binding, zinc ion binding, methyltransferase activity, histone-lysine N-methyltransferase activity (likely a histone H3 methyltransferase)	AF334407	N/A	GT/AG
25	NM_145019*	+/-	3/5	108(36)	582	675	LTR9	ERV1[1]	hypothetical protein FLJ30707 (*FLJ30707*)	structural constituent of ribosome	AK096364	DA745625	* donor (c & h & r = GT); acceptor (c & h = AG, r = deletion)

Table A.1 continued

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions [7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
26	NM_001014830	-/+	3/5	75(25)	324	327	MLT1A0	MaLR[1]	hypothetical protein LOC196913 (*LOC196913*)	undefined	AF390030	DA935043, DA947999	GT/AG
27	NM_033141*	-/+	11/13	42(14)	1118	589	MLT2B3	ERVL[1]	mitogen-activated protein kinase kinase kinase 9 (*MAP3K9*)	ATP binding, nucleotide binding, transferase activity, MAP kinase kinase activity, protein-tyrosine kinase activity, JUN kinase kinase kinase activity, protein homodimerization activity, protein serine/ threonine kinase activity	AF251442, BX648924	N/A	GT/AG
28	NM_007319*	+/-	8/11*	92	374	1701	MER52A	ERV1[1]	presenilin 1 (Alzheimer disease 3) (*PSEN1*), transcript variant I-374	protein binding, Notch receptor processing, amyloid precursor protein catabolism (regulation of amyloid precursor protein (APP) processing through their effects on a gamma-secretase, an enzyme that cleaves APP; involvement in the cleavage of the Notch receptor)	U40380, AF416717	N/A	no orthologous exonic sequences in chimpanzee and rhesus monkey
29	NM_020552*	+/-	2/5	67	105	129	MLT1D	MaLR[1]	T-cell leukemia/ lymphoma 6 (*TCL6*), transcript variant TCL6b1	unknown (a candidate gene for leukemogenesis)	AB035335	BX390485	donor (c & h = GT, r = GA); acceptor (all = AG)
30	NM_020553*	+/-	5/8	67	119	129	MLT1D	MaLR[1]	T-cell leukemia/ lymphoma 6 (*TCL6*), transcript variant TCL6c1	unknown (a candidate gene for leukemogenesis)	AB035337	BX390485	donor (c & h = GT, r = GA); acceptor (all = AG)

Table A.1 continued

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions [7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
31	NM_020554*	+/-	5/8	67	163	129	MLT1D	MaLR[1]	T-cell leukemia/ lymphoma 6 (*TCL6*), transcript variant TCL6d1	unknown (a candidate gene for leukemogenesis)	AB035338	BX390485	donor(c & h = GT, r = GA); acceptor (all = AG)
32	NM_005624*	+/-	4/5	120(40)	150	390	MLT1K	MaLR[1]	chemokine (C-C motif) ligand 25 (*CCL25*)	hormone activity, chemokine activity, chemotaxis, sensory perception, inflammatory response, G-protein coupled receptor protein signaling pathway	U86358	BX106823, AA295945	GT/AG
33	NM_005867*	-/-	2/3	102(34)	118	364	MLT2C1	ERVL[1]	Down syndrome critical region gene 4 (*DSCR4*)	unknown (contribution to the pathogenesis of many characteristics of Down syndrome, including morphological features, hypotonia, and mental retardation)	AB000099, BC069729, BC096162, BC096163, BC096164	CB993317, AW664531, BX112350, CD243838	GT/AG
34	NM_003159*	+/+	19/21*	84(28)	1030	398	MLT1J	MaLR[1]	cyclin-dependent kinase-like 5 (*CDKL5*), transcript variant I	ATP binding, nucleotide binding, transferase activity, protein-tyrosine kinase activity, protein serine/ threonine kinase activity, protein amino acid phosphorylation (association with X-linked infantile spasm syndrome (ISSX))	AY217744, Y15057, BC036091	N/A	no orthologous exonic sequence in chimpanzee
35	NM_024332*	+/-	8/12*	75(25)	316	528	MLT1J	MaLR[1]	chromosome X open reading frame 53 (*CXorf53*), transcript variant 1	unknown	BC002999, BC006540, AY438030, S68015	BU570969	no orthologous exonic sequences in chimpanzee and rhesus monkey
36	NM_207358*	-/-	4/10	118	139	1711	HERV16	ERVL[2]	hypothetical protein LOC339789 (*LOC339789*)	undefined	BC043563, AK127578	BI911310	donor (all = GT); acceptor (c & h = AG, r = AT)
37	NM_001004352*	+/+	4/7	187	141	5613	HERVL-A2	ERVL[2]	FLJ16323 protein (*FLJ16323*)	hydrolase activity, dUTP metabolism, nucleotide metabolism	AK131322	N/A	no orthologous exonic sequence in chimpanzee

Table A.1 continued

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions [7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
38	NM_004052*	-/+	3/6	85	194	87	HERV70	ERV1[2]	BCL2/adenovirus E1B 19kDa interacting protein 3 (*BNIP3*), nuclear gene encoding mitochondrial protein	induce apoptosis, anti-apoptosis, defense response to virus	BC009342, CR626676, BC021989, BC009342, AF002697, AK222626, BC067818, BC080643	CR995431, BX377664, BX404184, BX361486, BX361092, BX402244, BM799844, BM906527, BX445440, BM460721	GT/AG (GT position is not covered by HERV70)
39	NM_004052*	-/+	2/6	151	194	154	HERV70	ERV1[2]	BCL2/adenovirus E1B 19kDa interacting protein 3 (*BNIP3*), nuclear gene encoding mitochondrial protein	induce apoptosis, anti-apoptosis, defense response to virus	BC009342, CR626676, BC021989, BC009342, AF002697, AK222626, BC067818, BC080643	CR995431, BX377664, BX404184, BX361486, BX361092, BX402244, BM799844, BM906527, BM460721, BX333924	GT/AG (GT position is not covered by HERV70)
40	NM_014317*	+/+	2/12	33(11)	415	1493	THE1D-int	MaLR[2]	prenyl (decaprenyl) diphosphate synthase, subunit 1 (*PDSS1*)	transferase activity, isoprenoid biosynthesis	AK024802, AK223414, CR591586, AB210838, CR609440, BC063635, BC049211	DA565436, BX404180, CB996378, AL558274, CN297657, CN297656, AL560834, BF982221, DA969880, BI761095, BI916825, BQ212558, DR155397, CB152847	* GT/AG

Table A.1 continued

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions [7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
41	NM_080661*	+/+	2/7	124	333	701	MER57B-int	ERV1[2]	glycine-N-acyltransferase-like 1 (*GLYATL1*)	acyltransferase activity, transferase acivity	BC008353, AK091965, AX747281	DA082180, T80698, DA355423, BM924769, DA631846, BE265837, CA397661, CF122488	no orthologous exonic sequence in rhesus monkey
42	NM_016488*	+/+	12/13*	225(75)	458	2904	HERVK22	ERVK[2]	periphilin 1 (*PPHLN1*), transcript variant 1	keratinization (may play a role in epithelial differentiation and contribute to epidermal integrity and barrier formation)	AY039238	N/A	* GT/AG
43	NM_015138*	+/+	14/18	58	585	85	HERVIP10F	ERV1[2]	Rtf1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae) (*RTF1*)	undefined	D87440, BC015052	CX783855, BQ421107, CN334996, DB063883, CB989235, BF795182, CN334990, AU127092, BQ230838, BQ354365	GT/AG (T position is not covered by HERVIP10F)
44	NM_015547*	+/+	16/17*	153	607	3404	MER9, HERVK9	ERVK[3]	acyl-CoA thioesterase 11 (*ACOT11*), transcript variant 1	acyl-CoA thioesterase activity, hydrolase activity, serine esterase activity	AB014607, AF416921	N/A	* GT/AG
45	NM_016488*	+/+	11/13*	101	458	1356	LTR22A, HERVK22	ERVK[3]	periphilin 1 (*PPHLN1*), transcript variant 1	keratinization (may play a role in epithelial differentiation and contribute to epidermal integrity and barrier formation)	AY039238, BC025306	BI258755, BM838224	* GT/AG
first CDS exon													
1	NM_144663*	-/-	1/4	141(47)	217	580	LTR18B	ERVL[1]	chromosome 11 open reading frame 40 (*C11orf40*)	undefined	AF439154	N/A	* donor (c & h = GT, r = GA), no flanking acceptor: first exon of gene

Table A.1 continued

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions [7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
2	NM_001012274*	+/+	2/3	106	117	998	MER61A, MER61A-int	ERV1[3]	chromosome 1 open reading frame 99 (*C1orf99*)	undefined	BC040856	BM452121	* GT/AG
last CDS exon													
1	NM_013329*	-/-	16/16*	114(38)	815	631	MER39B	ERV1[1]	chromosome 21 open reading frame 66 (*C21orf66*), transcript variant 2	DNA binding, regulation of DNA-dependent transcription	AY033904	N/A	* acceptor (c & h & r = AG), no flanking donor: last exon of gene
2	NM_015845*	-/+	14/14*	190	586	656	MLT2F	ERVL[1]	methyl-CpG binding domain protein 1 (*MBD1*), transcript variant 2	metal ion binding, methyl-CpG binding, transcription corepressor activity	AF078831	N/A	acceptor(c & h & r = AG), no flanking donor: last exon of gene
3	NM_181538*	-/-	2/2	59	279	491	LTR72B	ERV1[1]	gap junction protein, epsilon 1, 29kDa (*GJE1*)	connexon channel activity	AF503615, AY297109	N/A	* acceptor (c & h & r = AG), no flanking donor: last exon of gene
4	NM_001015884	-/-	4/4	33(11)	115	384	MSTA	MaLR[1]	RPB11b2alpha protein (*POLR2J3*)	undefined	AJ277741	BI223488, BE936699, DW433233	* acceptor (c & h & r = AG), no flanking donor: last exon of gene
single protein coding exon													
1	NM_001007236	-/-	1/1	5640 (1880)	1879	7536	HERVK	ERVK[2]	endogenous retroviral sequence K, 6 (*ERVK6*)	aspartic-type endopeptidase activity, metal ion binding, nucleic acid binding, peptidase activity, zinc ion binding, DNA transposition, proteolysis, virus life cycle	Partial CDS: AF080233, DQ069911, DQ069913, U87590, U87591, U87592	N/A	no orthologous exonic sequence in chimpanzee and rhesus monkey
2	NM_001007236 (different genomic location from the previous entry)	-/-	1/1	5640 (1880)	1879	7535	HERVK	ERVK[2]	endogenous retroviral sequence K, 6 (*ERVK6*)	aspartic-type endopeptidase activity, metal ion binding, nucleic acid binding, peptidase activity, zinc ion binding, DNA transposition, proteolysis, virus life cycle	Partial CDS: AF080233, DQ069911, DQ069913, U87590, U87591, U87592	N/A	no orthologous exonic sequence in chimpanzee and rhesus monkey

Table A.1 continued

	RefSeq gene[1]	strand (gene/ LRTS)[2]	exon no/ total exon[3]	exon length, nt[4]	protein length, aa	LRTS length, nt[5]	LRTS sub-family	LRTS class/ family[6]	Gene annotation	GO descriptions[7]	GenBank mRNAs with exonized LRTS[8]	ESTs with exonized LRTS[9]	sequences in place of splice sites flanking orthologous exons[10]
3	NM_001007253	-/-	1/1	1696	564	8428	HERV3	ERV1[2]	endogenous retroviral unknown sequence 3 (includes zinc finger protein H-plk/HPF9) (*ERV3*)		N/A	N/A	no orthologous exonic sequence in chimpanzee

(1) RefSeq gene accession number. The entries marked with arterisk are known protein coding genes with the products listed in SWISS-PROT, TrEMBL, and TrEMBL-NEW and their corresponding mRNAs presented in the GenBank records. The adjacent entries in gray shading indicate the redundant exons, exons with exact or overlap boundaries participated in different alternative transcripts (derived from the same LRTS). Item 10-14 of internal protein coding exons could be the result of duplication of genes or gene regions, rather than direct LTR element insertion.
(2) Orientation of gene and LRTS assigned as gene strand/ LRTS strand. +, sense strand; -, antisense strand
(3) The position of exon on a gene/ total number of exons. The arterisk indicates the alternatively spliced exon (exons which are not present in all transcript variants).
(4) The length of LRTS-derived exon. In case of length divisible by three, the number in parenthesis shows the length of additional amino acid sequence provided by LRTS.
(5) The length of LRTS fragment covering the exon.
(6) Class/ family of LRTS. The superscript number refers to the part of LRTS covering an exon. 1, only LTR; 2, only internal sequence (sequence between flanking LTRs containing traditional viral genes); 3, both LTR and internal sequence.
(7) Gene descriptions according to the Gene Ontology (GO). undefined: no match for the gene in GO. unknown: gene product whose process, function, or localization is not known or cannot be inferred.
(8) GenBank mRNAs and (9) ESTs confirming the existence of LRTS-derived exon in human. Data were derived from the UCSC genome browser and the Entrez Gene.

(10) Sequences in place of splice sites flanking the LRTS-derived exon compared between human (May 2004, hg17), chimpanzee (Nov 2003, panTro1) and rhesus monkey (Jan 2006, rheMac2). h, human; c, chimpanzee; r, rhesus monkey. The data were derived from the multiple sequence alignment of the target exon of human and homologous sequences of chimpanzee and rhesus monkey retrieved from the UCSC genome browser. The entries marked with arterisk suggest possible primate-specific LTR element insertion (the homologous LRTS regions are present in chimpanzee and rhesus monkey but not in other vertebrates; mouse, rat, rabbit, dog, cow, armadillo, elephant, tenrec, opossum, chicken, frog, zebrafish, tetraodon, fugu).

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

Table B.1: List of 124 TE-associated Pfam protein domains

Accession	ID	Description
PF00075	RnaseH	RNase H
PF00077	RVP	Retroviral aspartyl protease
PF00078	RVT_1	Reverse transcriptase (RNA-dependent DNA polymerase)
PF00098	zf-CCHC	Zinc knuckle
PF00424	REV	REV protein (anti-repression trans-activator protein)
PF00429	TLV_coat	ENV polyprotein (coat polyprotein)
PF00469	F-protein	Negative factor, (F-Protein) or Nef
PF00516	GP120	Envelope glycoprotein GP120
PF00517	GP41	Envelope Polyprotein GP41
PF00522	VPR	VPR/VPX protein
PF00539	Tat	Transactivating regulatory protein (Tat)
PF00540	Gag_p17	gag gene protein p17 (matrix protein)
PF00552	Integrase	Integrase DNA binding domain
PF00558	Vpu	Vpu protein
PF00559	Vif	Retroviral Vif (Viral infectivity) protein
PF00589	Phge_integrase	Phage integrase family
PF00607	Gag_p24	gag gene protein p24 (core nucleocapsid protein)
PF00665	rve	Integrase core domain
PF00692	dUTPase	dUTPase
PF00872	Transposase_mut	Transposase, Mutator family
PF00906	Hepatitis_core	Hepatitis core antigen
PF00971	EIAV_GP90	EIAV coat protein, gp90
PF00979	Reovirus_cap	Reovirus outer capsid protein, Sigma 3
PF01021	TYA	TYA transposon protein
PF01045	EIAV_GP45	EIAV glycoprotein, gp45
PF01054	MMTV_SAg	Mouse mammary tumour virus superantigen
PF01140	Gag_MA	Matrix protein (MA), p15
PF01141	Gag_p12	Gag polyprotein, inner coat protein p12
PF01359	Transposase_1	Transposase
PF01385	Transposase_2	Probable transposase
PF01498	Transposase_5	Transposase
PF01526	Transposase_7	Transposase

Table B.1 continued

Accession	ID	Description
PF01527	Transposase_8	Transposase
PF01548	Transposase_9	Transposase
PF01609	Transposase_11	Transposase DDE domain
PF01610	Transposase_12	Transposase
PF01695	IstB	IstB-like ATP binding protein
PF01710	Transposase_14	Transposase
PF01797	Transposase_17	Transposase IS200 like
PF02022	Integrase_Zn	Integrase Zinc binding domain
PF02093	Gag_p30	Gag P30 core shell protein
PF02228	Gag_p19	Major core protein p19
PF02281	Transposase_Tn5	Transposase Tn5 dimerisation domain
PF02316	Mu_DNA_bind	Mu DNA-binding domain
PF02337	Gag_p10	Retroviral GAG p10 protein
PF02371	Transposase_20	Transposase IS116/IS110/IS902 family
PF02411	MerT	MerT mercuric transport protein
PF02720	DUF222	Domain of unknown function DUF222
PF02813	Retro_M	Retroviral M domain
PF02892	zf-BED	BED zinc finger
PF02914	Mu_transposase	Bacteriophage Mu transposase
PF02920	Integrase_DNA	DNA binding domain of tn916 integrase
PF02959	Tax	HTLV Tax
PF02992	Transposase_21	Transposase family tnp2
PF02994	Transposase_22	L1 transposable element
PF02998	Lentiviral_Tat	Lentiviral Tat protein
PF03004	Transposase_24	Plant transposase (Ptta/En/Spm family)
PF03017	Transposase_23	TNP1/EN/SPM transposase
PF03050	Transposase_25	Transposase IS66 family
PF03056	GP36	Env gp36 protein (HERV/MMTV type)
PF03078	ATHILA	ATHILA ORF-1 family
PF03108	MuDR	MuDR family transposase
PF03184	DDE	DDE superfamily endonuclease
PF03221	Transposase_Tc5	Tc5 transposase
PF03274	Foamy_BEL	Foamy virus BEL 1/2 protein
PF03276	Gag_spuma	Spumavirus gag protein
PF03400	Transposase_27	IS1 transposase
PF03408	Foamy_virus_ENV	Foamy virus envelope protein
PF03539	Spuma_A9PTase	Spumavirus aspartic protease (A9)
PF03708	Avian_gp85	Avian retrovirus envelope protein, gp85
PF03716	WCCH	WCCH motif
PF03732	Retrotrans_gag	Retrotransposon gag protein
PF03811	Ins_element1	Insertion element protein
PF04094	DUF390	Protein of unknown function (DUF390)
PF04160	Borrelia_orfX	Orf-X protein

Table B.1 continued

Accession	ID	Description
PF04195	Transposase_28	Putative gypsy type transposon
PF04218	CENP-B_N	CENP-B N-terminal DNA-binding domain
PF04236	Transp_Tc5_C	Tc5 transposase C-terminal domain
PF04582	Reo_sigmaC	Reovirus sigma C capsid protein
PF04693	Transposase_29	Archaeal putative transposase ISC1217
PF04740	Transposase_30	Bacillus transposase protein
PF04754	Transposase_31	Putative transposase, YhgA-like
PF04827	Plant_tran	Plant transposon protein
PF04937	DUF659	Protein of unknown function (DUF 659)
PF04986	Transposase_32	Putative transposase
PF05052	MerE	MerE protein
PF05344	DUF746	Domain of Unknown Function (DUF746)
PF05380	Peptidase_A17	Pao retrotransposon peptidase
PF05399	EVI2A	Ectropic viral integration site 2A protein (EVI2A)
PF05457	Transposase_33	Sulfolobus transposase
PF05485	THAP	THAP domain
PF05598	DUF772	Sulfolobus solfataricus protein of unknown function (DUF772)
PF05599	Deltaretro_Tax	Deltaretrovirus Tax protein
PF05621	TniB	Bacterial TniB protein
PF05699	hATC	hAT family dimerisation domain
PF05717	Transposase_34	IS66 Orf2 like protein
PF05754	DUF834	Domain of unknown function (DUF834)
PF05840	Phage_GPA	Bacteriophage replication gene A protein (GPA)
PF05851	Lentivirus_VIF	Lentivirus virion infectivity factor (VIF)
PF05858	BIV_Env	Bovine immunodeficiency virus surface envelope protein (ENV)
PF05928	Zea_mays_MuDR	Zea mays MURB-like protein (MuDR)
PF06527	TniQ	TniQ
PF06815	RVT_connect	Reverse transcriptase connection domain
PF06817	RVT_thumb	Reverse transcriptase thumb domain
PF07253	Gypsy	Gypsy protein
PF07282	Transposase_35	Putative transposase DNA-binding domain
PF07567	zf-C2HC_plant	Protein of unknown function, DUF1544
PF07572	BCNT	Bucentaur or craniofacial development
PF07592	Transposase_36	Rhodopirellula transposase
PF07727	RVT_2	Reverse transcriptase (RNA-dependent DNA polymerase)
PF07999	RHSP	Retrotransposon hot spot protein
PF08284	RVP_2	Retroviral aspartyl protease
PF08333	DUF1725	Protein of unknown function (DUF1725)
PF08483	IstB_N	IstB-like ATP binding N-terminal
PF08705	Gag_p6	Gag protein p6

Table B.1 continued

Accession	ID	Description
PF08721	TnsA_C	TnsA endonuclease C terminal
PF08722	TnsA_N	TnsA endonuclease N terminal
PF08723	Gag_p15	Gag protein p15
PF09035	Tn916-Xis	Excisionase from transposon Tn916
PF09039	Mu_I-gamma	Mu DNA binding, I gamma subdomain
PF09077	Phage-MuB_C	Mu B transposition protein, C terminal
PF09293	RNaseH_C	T4 RNase H, C terminal
PF09299	Mu-transpos_C	Mu transposase, C-terminal
PF09322	DUF1979	Domain of unknown function (DUF1979)

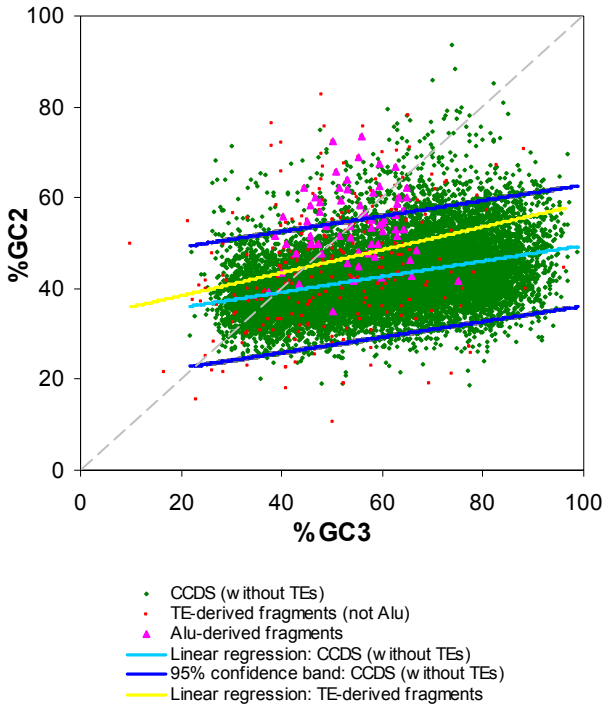

Figure B.1: The GC composition of Alu-derived gene fragments. Scatter plots of %G+C of second (GC2) versus third (GC3) codon positions for Alu-derived gene fragments (pink), non Alu TE-derived gene fragments (red) and non TE-associated genes (green) are shown along with linear regression trends and confidence interval.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

Table C.1: mRNAs anticorrelated with hsa-mir-130b and their associated GO terms

Table C.1 continued

Name[a]	GO[b]	*P*-value[c]
A4GNT	GO:0044237 cellular metabolism	0.0037
BACH2	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
BRPF1	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
CAPN10	GO:0044237 cellular metabolism	0.0037
CCNB1	GO:0050789 regulation of biological process	0.0017
CCRN4L	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
COX7A2L	GO:0044237 cellular metabolism	0.0037
EEF1B2	GO:0044237 cellular metabolism	0.0037
EPM2A	GO:0031323 regulation of cellular metabolism	0.0008
ERBB4	GO:0044237 cellular metabolism	0.0037
EREG	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
GADD45A	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
GALR1	GO:0050789 regulation of biological process	0.0017
GTF2H1	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
HOXD1	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
JARID2	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
KLF13	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
MAX	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
MBD4	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
MITF	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
MRPS30	GO:0044237 cellular metabolism	0.0037
MYB	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05

Table C.1 continued

Name[a]	GO[b]	*P*-value[c]
NOX1	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
PITPNM2	GO:0008152 metabolism	0.0068
PPIG	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
PRKAA2	GO:0044237 cellular metabolism	0.0037
PRMT7	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
PSMA8	GO:0044237 cellular metabolism	0.0037
PTGER3	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
PTH	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
RPS2	GO:0044237 cellular metabolism	0.0037
SIRT6	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
SIRT7	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
TGIF2	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
TP73L	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
UBE2W	GO:0044237 cellular metabolism	0.0037
VGLL2	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
WHSC1L1	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05
ZNF430	GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.96E-05

[a]Gene name for the anticorrelated mRNA
[b]Over-represented biological process GO term and description
[c]*P*-value associated with that GO term

Table C.2: Conserved RNA secondary structures that co-locate with human TE sequences

Name[a]	Coords[b]	TE[c]
6597_0_+_74	chr1:6483733-6483759(+)	Charlie8
15086_0_-_78	chr1:15041842-15041859(-)	HAL1
24981_0_+_74	chr1:23415863-23415881(+)	MIRb

Table C.2 continued

Name[a]	Coords[b]	TE[c]
25288_0_-_83	chr1:23621848-23621877(-)	MIRb
25923_0_+_48	chr1:24115467-24115495(+)	MIR3
30647_0_+_38	chr1:27752374-27752433(+)	MIRb
30820.5_0_-_68	chr1:27791805-27791823(-)	MIR3
38333_0_+_77	chr1:34707582-34707607(+)	Charlie7
43475_0_+_88	chr1:38294035-38294051(+)	L3
46777_0_+_71	chr1:40440123-40440153(+)	MER103
48001_0_+_47	chr1:41214471-41214487(+)	L2
48998_0_-_51	chr1:41888110-41888146(-)	L3
51113_0_+_73	chr1:43734410-43734424(+)	L1M5
58555_0_+_65	chr1:49094515-49094531(+)	L1ME4a
60187_0_+_106	chr1:50667381-50667396(+)	L1ME4a
72612_0_-_79	chr1:61379203-61379221(-)	L3
73590_0_-_75	chr1:61923623-61923650(-)	Charlie2
82264_0_-_68	chr1:71707409-71707427(-)	L3
85615_0_+_83	chr1:76474930-76474947(+)	MIRb
92562_0_-_56	chr1:83112753-83112791(-)	L2
108307_0_+_78	chr1:97287966-97287983(+)	L3b
112207_0_+_95	chr1:101216647-101216665(+)	MIR3
120809_0_+_79	chr1:111021701-111021719(+)	MIR
122080_0_-_62	chr1:112177611-112177631(-)	MIR
124780_0_-_66	chr1:114214379-114214407(-)	MIRb
132990_0_-_73	chr1:142748623-142748637(-)	MIR3
148967_0_+_106	chr1:157421181-157421197(+)	L2
149463_0_+_61	chr1:157999850-157999872(+)	L4
154215_0_+_65	chr1:162123173-162123189(+)	Tigger8
161465_0_+_90	chr1:169013965-169013984(+)	MER45A
177588_0_-_76	chr1:182715187-182715207(-)	MIRb
184070_0_-_57	chr1:192397663-192397683(-)	MER90
188643_0_+_89	chr1:198985820-198985838(+)	L2
190021_0_+_47	chr1:199883257-199883273(+)	MIRb
199932_0_-_79	chr1:206789414-206789432(-)	Charlie2
228731_0_+_62	chr1:242250135-242250150(+)	MER53
230542_0_-_67	chr1:244286075-244286098(-)	L1MB3
1190052_0_-_69	chr2:26795482-26795520(-)	L1ME4a
1194893_0_+_73	chr2:29846946-29846960(+)	HAL1
1201447_0_+_65	chr2:37374164-37374180(+)	L4
1208495_0_+_83	chr2:44171423-44171440(+)	LTR33
1208526_0_-_109	chr2:44198198-44198220(-)	MLT1J
1222791_0_+_110	chr2:59759711-59759730(+)	L1M5
1228152_0_-_94	chr2:64565601-64565616(-)	MIR3
1229575_0_+_76	chr2:65898157-65898173(+)	LTR33
1239485_0_+_80	chr2:75090614-75090638(+)	L1ME4a

Table C.2 continued

Name[a]	Coords[b]	TE[c]
1239721_0_+_79	chr2:75407124-75407142(+)	MARNA
1239765_0_+_119	chr2:75520847-75520862(+)	L1MC
1243784_0_-_105	chr2:81417721-81417741(-)	MIRb
1245396_0_-_65	chr2:85055655-85055671(-)	MER58B
1247043_0_+_65	chr2:86563203-86563222(+)	L3
1250762_0_-_79	chr2:95726497-95726515(-)	L2
1257853_0_+_43	chr2:103875041-103875063(+)	MIRm
1258738_0_-_45	chr2:104657492-104657535(-)	L1ME4a
1265559_0_+_78	chr2:113813281-113813307(+)	L2
1280461_0_+_93	chr2:136964701-136964729(+)	Tigger8
1288189_0_+_138	chr2:145136397-145136417(+)	L1MC5
1304651_0_+_81	chr2:161680961-161680976(+)	MIRb
1307871_0_+_94	chr2:164440131-164440148(+)	L3b
1312547_0_+_65	chr2:169163008-169163024(+)	Kanga2_a
1318523_0_+_114	chr2:174053048-174053068(+)	L2
1319174_0_-_68	chr2:174641664-174641691(-)	L1MC4a
1320276_0_-_65	chr2:175480749-175480768(-)	L1ME4a
1344295_0_+_73	chr2:200120342-200120356(+)	MIRm
1346880_0_-_90	chr2:202119446-202119466(-)	L1MB8
1353436_0_-_43	chr2:206288743-206288763(-)	MIRb
1363676_0_-_76	chr2:215455658-215455686(-)	MARNA
1365045_0_+_67	chr2:216418189-216418209(+)	MLT1K
1371955_0_+_109	chr2:220595160-220595181(+)	Arthur1
1383146_0_-_65	chr2:232693947-232693963(-)	L2
1389370_0_-_70	chr2:240141186-240141212(-)	L1M5
1521839_0_+_74	chr3:10540626-10540652(+)	Charlie8
1525689_0_+_100	chr3:14249470-14249491(+)	L3
1530481_0_+_59	chr3:18526135-18526168(+)	L2
1542101_0_+_74	chr3:35354374-35354400(+)	L1ME3A
1558705_0_+_79	chr3:50643228-50643246(+)	MIRb
1572162_0_-_74	chr3:60630961-60630979(-)	L1M5
1584630_0_-_110	chr3:70684589-70684609(-)	L1ME3A
1590226_0_-_100	chr3:74756051-74756070(-)	MER45B
1593242_0_-_80	chr3:78065198-78065217(-)	MIRb
1595751_0_+_87	chr3:81453304-81453326(+)	MIRb
1605591_0_+_81	chr3:100984579-100984594(+)	L3b
1619386_0_-_86	chr3:115872816-115872837(-)	MER69A
1624711_0_+_126	chr3:120022139-120022157(+)	L1ME4a
1638166_0_-_65	chr3:133240374-133240390(-)	MLT1J
1651552_0_-_59	chr3:145681874-145681905(-)	L2
1664761_0_+_78	chr3:161388886-161388903(+)	MER3
1665104_0_-_61	chr3:161708071-161708106(-)	L2
1669125_0_+_65	chr3:169470145-169470170(+)	L1ME4a

Table C.2 continued

Name[a]	Coords[b]	TE[c]
1692753_0_+_65	chr3:194760060-194760076(+)	MIRb
1705626_0_-_63	chr4:12690588-12690606(-)	L2
1709742_0_-_82	chr4:17391442-17391474(-)	L2
1714124_0_-_124	chr4:23001557-23001573(-)	L2
1719629_0_+_105	chr4:27859462-27859480(+)	LTR33
1726312_0_-_63	chr4:39411734-39411760(-)	MIRb
1730972_0_-_56	chr4:46681709-46681733(-)	L1ME3B
1747758_0_-_63	chr4:74275595-74275629(-)	L1M5
1753586_0_-_83	chr4:81294838-81294860(-)	MIRb
1754111_0_+_88	chr4:81949213-81949228(+)	L3
1764240_0_-_76	chr4:95780967-95780983(-)	MIR3
1768888_0_-_54	chr4:101481090-101481113(-)	LTR33
1772239_0_+_73	chr4:106784430-106784444(+)	MIR3
1773181_0_+_100	chr4:108002087-108002108(+)	LTR68
1813460_0_-_96	chr4:158919215-158919239(-)	MER5A
1827139_0_+_83	chr4:181230274-181230297(+)	MLT1C
1827751_0_+_75	chr4:181988895-181988914(+)	MIRb
1829210_0_-_55	chr4:183054460-183054490(-)	MIR3
1830949_0_+_65	chr4:184054997-184055013(+)	MIRm
1842241_0_+_61	chr5:9629511-9629533(+)	Charlie2
1860397_0_+_94	chr5:38819321-38819337(+)	MIR3
1862891_0_+_55	chr5:42157514-42157542(+)	LTR16A1
1868925_0_+_75	chr5:53320758-53320785(+)	MIRb
1874290_0_-_110	chr5:59086690-59086709(-)	MER113
1876462_0_-_84	chr5:61101641-61101659(-)	MIRb
1900649_0_+_81	chr5:88600908-88600928(+)	MARNA
1904329_0_+_85	chr5:92449669-92449694(+)	L3
1912951_0_+_71	chr5:103360475-103360491(+)	MIRb
1919694_0_-_71	chr5:112530605-112530621(-)	L3b
1920501_0_+_72	chr5:113735156-113735173(+)	L2
1929047_0_+_75	chr5:124046980-124047023(+)	HAL1
1944753_0_+_72	chr5:139204106-139204134(+)	Charlie7
1944754_0_-_54	chr5:139204158-139204181(-)	Charlie7
1949916_0_-_74	chr5:141188281-141188299(-)	L1MC
1951254_0_-_65	chr5:142156112-142156128(-)	MIR3
1959488_0_-_87	chr5:149669722-149669736(-)	L2
1966281_0_+_83	chr5:156681824-156681841(+)	MIR3
1967253_0_+_52	chr5:157463563-157463620(+)	L1ME4a
1969951_0_+_79	chr5:159059780-159059812(+)	L4
1970767_0_+_67	chr5:159755119-159755139(+)	Charlie11
1973974_0_-_68	chr5:163838798-163838819(-)	L2
1980933_0_+_81	chr5:168747748-168747763(+)	Charlie1a
1987527_0_+_59	chr5:175727565-175727628(+)	L2

Table C.2 continued

Name[a]	Coords[b]	TE[c]
1988415_0_+_77	chr5:176255041-176255066(+)	MIRb
1874239_0_+_93	chr5:59035208-59035277(+)	L3
1999329_0_+_67	chr6:7450078-7450092(+)	MIR3
2006707_0_+_93	chr6:15110399-15110413(+)	MIRb
2012915_0_+_82	chr6:21665770-21665786(+)	Charlie8
2020947_0_+_75	chr6:31654313-31654336(+)	LTR42
2029187_0_+_46	chr6:37530191-37530227(+)	L3b
2029634_0_-_100	chr6:37784559-37784580(-)	L1ME4a
2032923_0_+_87	chr6:40857398-40857412(+)	L1ME4a
2037716_0_-_160	chr6:44077681-44077715(-)	MER53
2047183_0_+_78	chr6:54681376-54681393(+)	L2
2075048_0_-_91	chr6:94484941-94484963(-)	ERVL-E
2078828_0_-_111	chr6:99509678-99509696(-)	MIRb
2085959_0_+_73	chr6:107897109-107897123(+)	MER5B
2096533_0_+_62	chr6:119656124-119656165(+)	L3
2171717_0_+_110	chr7:34814854-34814873(+)	HAL1b
2177366_0_-_111	chr7:41288277-41288304(-)	L1ME4a
2195049_0_+_117	chr7:73161289-73161306(+)	MIR3
2203727_0_+_62	chr7:83830368-83830388(+)	L3
2214152_0_-_57	chr7:95381816-95381836(-)	MIR
2215010_0_-_104	chr7:95921415-95921438(-)	L1ME3
2216137_0_+_72	chr7:97120781-97120805(+)	MIR3
2229371_0_+_67	chr7:110280027-110280041(+)	LTR40a
2234441_0_+_52	chr7:115222951-115222981(+)	L2
2246941_0_+_73	chr7:128849628-128849649(+)	L2
2247695_0_+_119	chr7:129521829-129521855(+)	L1ME4a
2247695_1_+_65	chr7:129521966-129521985(+)	L1ME4a
2251058_0_-_95	chr7:131926833-131926869(-)	Charlie7
2277373_0_+_89	chr8:9268418-9268435(+)	L1MD2
2278817_0_+_62	chr8:10784215-10784248(+)	MIRb
2285985_0_+_74	chr8:20987404-20987426(+)	L2
2286909_0_+_87	chr8:21794450-21794464(+)	MIRb
2293775_0_+_77	chr8:28190662-28190683(+)	L1MC4a
2296966_0_-_112	chr8:31652907-31652946(-)	MIRb
2299343_0_+_68	chr8:33896886-33896904(+)	L1MC4
2299747_0_+_76	chr8:34293119-34293135(+)	MIRb
2301209_0_+_94	chr8:35801853-35801870(+)	L3
2306450_0_-_70	chr8:41798076-41798095(-)	L2
2317552_0_-_83	chr8:64247707-64247729(-)	MIRb
2322741_0_+_67	chr8:69743160-69743177(+)	L3
2330830_0_-_81	chr8:78984234-78984259(-)	MER5B
2337143_0_+_72	chr8:89013976-89013993(+)	L2
2339739_0_+_72	chr8:92822152-92822176(+)	L3

Table C.2 continued

Name[a]	Coords[b]	TE[c]
2339740_0_-_87	chr8:92822207-92822229(-)	L3
2348773_0_+_51	chr8:102229956-102230022(+)	Charlie9
2356873_0_+_100	chr8:110789279-110789300(+)	L3
2363824_0_-_100	chr8:119852398-119852414(-)	L4
2369557_0_+_74	chr8:125994973-125994991(+)	MIRb
2373181_0_+_76	chr8:130158434-130158450(+)	L2
2376487_0_+_73	chr8:134299648-134299669(+)	MIRb
2378900_0_-_81	chr8:138418970-138418990(-)	MER91B
2380602_0_-_97	chr8:141811863-141811931(-)	L2
2389613_0_+_87	chr9:3936317-3936331(+)	L2
2393693_0_+_78	chr9:8804094-8804111(+)	L2
2394223_0_+_76	chr9:9267671-9267699(+)	L1M4
2401146_0_-_96	chr9:16787222-16787246(-)	MIR
2401571_0_+_67	chr9:17060609-17060626(+)	L1ME4a
2412879_0_+_68	chr9:29844233-29844254(+)	L3
2421368_0_-_79	chr9:37811135-37811158(-)	L1MC4a
2426661_0_+_64	chr9:70297285-70297306(+)	MER91A
2431307_1_+_89	chr9:75353184-75353201(+)	L4
2442499_0_+_85	chr9:88825569-88825595(+)	L4
2447361_0_+_94	chr9:97429994-97430011(+)	MLT1K
2452438_0_-_120	chr9:101754094-101754108(-)	L4
2452597_0_+_91	chr9:101776189-101776211(+)	L4
2452602_0_-_68	chr9:101776449-101776467(-)	L4
2455634_0_-_64	chr9:105918396-105918420(-)	MER5A
2468629_0_+_104	chr9:117452580-117452603(+)	L2
2472869_0_+_76	chr9:121613122-121613138(+)	L2
2479521_0_+_75	chr9:126607543-126607574(+)	L1ME4a
2500550_0_-_83	chrX:10899595-10899617(-)	L4
2507070_0_-_85	chrX:15944486-15944505(-)	L1ME4a
2509218_0_-_95	chrX:17351726-17351746(-)	MIRb
2513321_0_+_81	chrX:19783471-19783486(+)	MER58A
2514009_0_+_54	chrX:20163518-20163543(+)	L1ME4a
2514011_0_+_147	chrX:20164132-20164161(+)	L1ME4a
2514641_0_+_84	chrX:20436611-20436629(+)	L2
2519737_0_+_67	chrX:24557155-24557175(+)	L1ME4a
2527121_0_+_47	chrX:32060447-32060480(+)	MIRb
2537738_0_-_61	chrX:43454292-43454314(-)	MIRb
2557941_0_+_75	chrX:71069003-71069018(+)	L2
2571697_0_+_73	chrX:97698707-97698728(+)	Charlie1a
2573092_0_-_86	chrX:99503268-99503289(-)	MIR
2576468_0_-_59	chrX:102471559-102471585(-)	L1MC4a
2583739_0_+_63	chrX:108529599-108529617(+)	L2
2584535_0_+_91	chrX:109185224-109185300(+)	MER91C

Table C.2 continued

Name[a]	Coords[b]	TE[c]
2585968_0_-_86	chrX:110177627-110177654(-)	L2
2598315_0_+_62	chrX:123586235-123586255(+)	MIRb
2604008_0_-_88	chrX:129506265-129506281(-)	HAL1b
2604832_0_-_51	chrX:129994594-129994630(-)	L3
2607024_0_-_68	chrX:131689852-131689873(-)	L1MB5
2613374_0_-_71	chrX:135974107-135974123(-)	MIR
2613853_0_+_60	chrX:136377304-136377323(+)	L1MC4
2625375_0_+_86	chrX:152562536-152562556(+)	L2
241830_0_-_89	chr10:13315027-13315044(-)	MIR
246588_0_-_100	chr10:18657192-18657208(-)	MER5B
251147_0_+_94	chr10:24324839-24324855(+)	MIRb
276291_0_+_66	chr10:62836157-62836220(+)	L1M5
278579_0_+_157	chr10:64671753-64671775(+)	L2
279306_0_-_71	chr10:65568421-65568437(-)	MLT1L
292265_0_+_95	chr10:77784926-77784944(+)	L2
295318_0_-_86	chr10:80133480-80133500(-)	L3b
296055_0_+_79	chr10:80679702-80679720(+)	MIRb
299725_0_+_88	chr10:86569717-86569742(+)	L1ME4a
333376_0_-_70	chr10:115808231-115808257(-)	MER46C
334961_0_+_78	chr10:117579937-117579954(+)	L2
341781_0_+_89	chr10:123290268-123290285(+)	L3
361933_0_+_88	chr11:6635631-6635654(+)	Kanga2_a
368338_0_-_100	chr11:11787492-11787508(-)	L3
370306_0_-_88	chr11:13416460-13416476(-)	L3
377681_0_+_96	chr11:19331037-19331062(+)	L3
394814_0_-_79	chr11:40886992-40887015(-)	L2
395263_0_+_46	chr11:41605707-41605732(+)	FordPrefect_a
397099_0_+_71	chr11:43869396-43869412(+)	MIRb
408823_0_+_73	chr11:59187569-59187590(+)	MIR
425555_0_+_71	chr11:71985685-71985701(+)	MIR
438440_0_-_74	chr11:83316345-83316367(-)	L2
438439_0_+_83	chr11:83316376-83316398(+)	L2
440997_0_+_81	chr11:85937263-85937278(+)	L4
444101_0_+_100	chr11:91430989-91431005(+)	MIR
466625_0_+_110	chr11:116956700-116956719(+)	MIR
471849_0_+_90	chr11:119880617-119880636(+)	L3
486187_0_+_68	chr11:130861130-130861151(+)	MIRb
492576_0_-_95	chr12:2125422-2125443(-)	MIRb
498153_0_+_67	chr12:6658215-6658235(+)	MIRm
522947_0_+_70	chr12:40826566-40826585(+)	L2
533638.0_0_-_122	chr12:50492331-50492353(-)	MIRb
542148_0_-_83	chr12:55246557-55246574(-)	LTR37B
569483_0_+_39	chr12:87985382-87985425(+)	L2

Table C.2 continued

Name[a]	Coords[b]	TE[c]
570380_0_-_86	chr12:88718407-88718434(-)	MIRb
571837_0_+_78	chr12:90583073-90583090(+)	MER58B
578709_0_+_108	chr12:96992583-96992607(+)	L1ME3B
637241_0_+_66	chr13:52383235-52383263(+)	L3b
645204_0_-_94	chr13:62341135-62341150(-)	MIR
648567_0_+_59	chr13:67872860-67872891(+)	MIRb
650887_0_+_68	chr13:71623788-71623824(+)	L3b
654127_0_+_82	chr13:74685170-74685186(+)	MER5A
710673_0_+_58	chr14:49857372-49857390(+)	MIR
731341_0_+_70	chr14:67762102-67762121(+)	L3b
737958_0_-_62	chr14:72903436-72903456(-)	MIRb
740946_0_-_72	chr14:74763371-74763399(-)	L2
746350_0_-_81	chr14:78670822-78670842(-)	MIRm
766865_0_+_57	chr14:100584767-100584820(+)	MER5A1
775713_0_+_77	chr15:25703141-25703162(+)	L1MCc
783373_0_+_94	chr15:33592451-33592467(+)	L1ME4a
795825_0_+_111	chr15:41875416-41875434(+)	MIRb
804788_0_-_75	chr15:50144808-50144835(-)	L1M1
823849_0_+_52	chr15:65945171-65945195(+)	L1ME4a
830936_0_-_60	chr15:71924252-71924271(-)	L4
842223_0_+_71	chr15:81565655-81565678(+)	MIR3
844067_0_+_100	chr15:83150156-83150175(+)	L1ME4a
851769_0_+_126	chr15:89250081-89250099(+)	L4
854345_0_+_71	chr15:91631904-91631924(+)	MIRb
857021_0_+_68	chr15:94257983-94258001(+)	MIRb
858814_3_-_112	chr15:95403622-95403655(-)	L4
872420_0_-_70	chr16:6175415-6175434(-)	MIRb
896537_0_+_81	chr16:30749660-30749680(+)	MIR
904577_0_+_91	chr16:49653282-49653304(+)	MER99
905411_0_+_74	chr16:50290185-50290231(+)	L3
909177_0_+_82	chr16:53170182-53170198(+)	MIRb
914771_0_+_81	chr16:57783182-57783208(+)	MIRb
918945_0_+_64	chr16:63576155-63576179(+)	L1M5
924692_0_-_70	chr16:67025377-67025409(-)	L1ME4a
928869_0_+_74	chr16:70304015-70304037(+)	MIR3
929116_0_-_165	chr16:70444938-70444954(-)	MIR
933812_0_-_89	chr16:74025554-74025571(-)	Tigger2
976169_0_+_86	chr17:24040248-24040268(+)	L1ME4a
989909_0_+_100	chr17:34009010-34009024(+)	MIR3
993737_0_+_67	chr17:36057125-36057139(+)	MIRb
1000039.8_0_+_109	chr17:39468501-39468532(+)	L1MC4
1015457_0_+_67	chr17:50377318-50377332(+)	MLT1C
1018093_0_+_74	chr17:52576474-52576492(+)	MIR3

Table C.2 continued

Name[a]	Coords[b]	TE[c]
1027140_0_-_109	chr17:58956205-58956226(-)	MER5A
1029871_0_+_79	chr17:60844235-60844253(+)	MIR
1044595_0_+_71	chr17:73614287-73614307(+)	L1MDa
1054900_0_-_75	chr18:6426001-6426016(-)	L1M5
1062229_0_+_73	chr18:18029737-18029751(+)	L1ME4a
1067853_0_+_86	chr18:22989505-22989533(+)	L3
1072891_0_+_61	chr18:29462119-29462141(+)	MIRb
1073425_0_+_104	chr18:30062939-30062963(+)	L2
1076640_0_-_90	chr18:33541894-33541913(-)	L2
1077026_0_-_100	chr18:33875539-33875554(-)	MIRb
1077028_0_-_58	chr18:33875730-33875789(-)	MIRb
1079884_0_-_53	chr18:37167167-37167185(-)	L3
1083125_0_+_82	chr18:41077120-41077141(+)	L2
1083586_0_-_71	chr18:41429464-41429484(-)	MIR
1084952_0_-_69	chr18:42863307-42863338(-)	L1ME4a
1085381_0_+_105	chr18:43238379-43238399(+)	MIR
1085682_0_-_76	chr18:43568621-43568637(-)	MER113
1100012_0_-_81	chr18:59852148-59852168(-)	L3
1105949_0_+_73	chr18:71397854-71397883(+)	L3b
1139206_0_-_87	chr19:37294434-37294456(-)	L2
1394080_0_-_52	chr20:2927811-2927835(-)	L1ME3B
1397189_0_+_110	chr20:6163591-6163611(+)	L2
1401791_0_+_119	chr20:10612937-10612952(+)	L2
1405784_0_-_100	chr20:14549750-14549770(-)	MIRb
1405783_1_+_46	chr20:14549818-14549865(+)	MIRb
1412553_0_+_87	chr20:20910851-20910873(+)	L1P4
1416308_0_-_95	chr20:29589818-29589838(-)	MIRb
1434998_0_-_100	chr20:44089910-44089928(-)	MIR
1435354_0_-_79	chr20:44235903-44235921(-)	MIR
1453725_0_+_67	chr21:14966802-14966816(+)	MIR3
1466070_0_-_70	chr21:33853177-33853203(-)	L2
1489729_0_+_59	chr22:28457760-28457800(+)	L1ME4a
1496941_0_+_79	chr22:35289947-35289989(+)	L1MC4
1498291_0_+_153	chr22:36325294-36325312(+)	L3b
1509614_0_-_55	chr22:48049650-48049680(-)	L1ME4a
3715_0_+_61	chr1:3131597-3131629(+)	MER121
52664_0_-_50	chr1:44571346-44571464(-)	Eulor9A
52940_0_-_77	chr1:44674828-44674849(-)	MER121
67625_0_+_68	chr1:57127369-57127387(+)	Eulor1
67626_0_-_76	chr1:57127400-57127465(-)	Eulor1
75315_0_+_61	chr1:63495303-63495320(+)	UCON8
82063_0_+_72	chr1:71463300-71463317(+)	MER133A
88341_0_+_95	chr1:79596295-79596315(+)	UCON15

Table C.2 continued

Name[a]	Coords[b]	TE[c]
88403_0_+_60	chr1:79660341-79660380(+)	UCON30
88624_0_-_75	chr1:80015139-80015158(-)	X6A_LINE
93362.2_0_-_68	chr1:83766400-83766418(-)	MER131
112743_0_+_87	chr1:101723969-101723998(+)	MER133A
130308_0_-_77	chr1:118651361-118651382(-)	MER121
130734_0_-_76	chr1:119254356-119254388(-)	MER121
154446_0_-_100	chr1:162492879-162492895(-)	UCON18
154818_0_-_64	chr1:162825371-162825437(-)	MER135
161005_0_+_64	chr1:168646670-168646694(+)	UCON26
187011_0_+_89	chr1:197245358-197245395(+)	MER121
188052_1_-_92	chr1:198460508-198460590(-)	Eulor3
198813_0_+_90	chr1:206237185-206237204(+)	MER136
204532_0_-_104	chr1:211522027-211522054(-)	UCON31
211312_0_-_76	chr1:217555499-217555519(-)	MER121
1171004_0_-_83	chr2:6568368-6568396(-)	MER121
1184389_0_+_73	chr2:22111720-22111741(+)	MER136
1221529_0_+_96	chr2:58895995-58896019(+)	X7C_LINE
1223513_0_+_97	chr2:60171167-60171200(+)	Eulor10
1231302_0_-_88	chr2:66979970-66979994(-)	Eulor3
1231553_0_+_75	chr2:67238894-67239028(+)	Eulor4
1241102_0_-_78	chr2:77353762-77353793(-)	MER123
1258257_0_+_85	chr2:104314401-104314489(+)	MER134
1258569_0_+_76	chr2:104524897-104524954(+)	Eulor1
1260469_0_+_45	chr2:107079289-107079332(+)	Eulor10
1260520_0_-_81	chr2:107258693-107258740(-)	UCON23
1285391_0_+_90	chr2:143702983-143703011(+)	MER136
1285391_1_+_129	chr2:143703042-143703075(+)	MER136
1285392_0_-_87	chr2:143703144-143703181(-)	MER136
1286513_0_+_57	chr2:144311903-144311946(+)	MER125
1288297_0_+_88	chr2:145186813-145186838(+)	Eulor4
1288651_0_+_86	chr2:145417357-145417378(+)	LmeSINE1b
1305099_0_+_74	chr2:161974382-161974408(+)	AmnSINE1_GG
1307386_0_-_85	chr2:164117565-164117584(-)	Eulor6D
1321435_0_+_100	chr2:176377724-176377743(+)	MER133A
1329277_0_+_76	chr2:181280870-181280886(+)	MER135
1343221_0_+_43	chr2:199281672-199281699(+)	X6A_LINE
1344357_0_+_95	chr2:200156156-200156174(+)	MER131
1344453_0_+_48	chr2:200237188-200237216(+)	MER121
1357076_0_-_85	chr2:208527969-208528001(-)	MER134
1361323_0_+_57	chr2:213067475-213067509(+)	Eulor5A
1372257_0_+_56	chr2:220785460-220785507(+)	UCON11
1373497_0_+_65	chr2:221709145-221709161(+)	MER121
1512259_0_+_84	chr3:886207-886231(+)	Eulor3

Table C.2 continued

Name[a]	Coords[b]	TE[c]
1529073_0_+_51	chr3:17512755-17512791(+)	UCON17
1537363_0_+_82	chr3:28861055-28861082(+)	Eulor4
1570585_0_+_100	chr3:59401515-59401529(+)	UCON17
1573547_0_+_44	chr3:61643441-61643518(+)	MER126
1573643_0_+_95	chr3:61718341-61718381(+)	MER134
1583919_0_+_136	chr3:70159597-70159621(+)	Eulor3
1584125_0_+_85	chr3:70282197-70282216(+)	MER121
1587903_0_+_63	chr3:72770417-72770435(+)	MER121
1589951_0_+_75	chr3:74512342-74512357(+)	MER129
1613192_0_-_38	chr3:109671038-109671090(-)	MER127
1619194_0_-_81	chr3:115772653-115772683(-)	UCON9
1620066_0_-_64	chr3:116298434-116298458(-)	Eulor1
1631238_0_-_62	chr3:125850420-125850440(-)	X7B_LINE
1644004_0_-_55	chr3:138650699-138650738(-)	UCON4
1651767_0_+_52	chr3:146074810-146074873(+)	Eulor3
1662217_0_+_62	chr3:159045808-159045847(+)	MER121
1668216_0_-_58	chr3:168436231-168436447(-)	MER126
1670278_0_-_120	chr3:170405686-170405715(-)	MER131
1678363_0_+_67	chr3:179445164-179445181(+)	X6B_LINE
1680024_0_-_54	chr3:181521306-181521359(-)	UCON29
1688359_0_+_58	chr3:189369310-189369328(+)	UCON7
1689205_1_+_86	chr3:190161865-190161886(+)	UCON7
1705501_0_+_62	chr4:12551247-12551267(+)	MER121
1705501_1_+_80	chr4:12551274-12551293(+)	MER121
1706954_0_-_78	chr4:14392438-14392473(-)	MER121
1710470_0_-_75	chr4:18287096-18287119(-)	X6B_LINE
1713985_0_+_114	chr4:22907781-22907802(+)	MER132
1713986_1_-_100	chr4:22907805-22907826(-)	MER132
1714957_0_+_71	chr4:23573735-23573758(+)	UCON2
1715903_0_+_88	chr4:24017990-24018005(+)	UCON2
1743298_0_-_89	chr4:67215793-67215838(-)	Eulor8
1743393_0_+_92	chr4:67394199-67394235(+)	Eulor5B
1757379_0_+_70	chr4:85466757-85466855(+)	MER134
1798767_0_+_50	chr4:142921647-142921690(+)	UCON9
1802422_0_-_75	chr4:147467073-147467100(-)	LF-SINE
1803203_0_+_84	chr4:148224139-148224169(+)	X7C_LINE
1828955_0_+_104	chr4:182906948-182906974(+)	Eulor6A
1829383_0_+_100	chr4:183149701-183149718(+)	LF-SINE
1830405_0_+_49	chr4:183690755-183690850(+)	MER135
1846135_0_+_73	chr5:15474550-15474564(+)	MER121
1850975_0_+_96	chr5:27214825-27214847(+)	Eulor3
1851065_0_+_102	chr5:27505032-27505076(+)	Eulor4
1864187_0_+_85	chr5:44037377-44037402(+)	MER121

Table C.2 continued

Name[a]	Coords[b]	TE[c]
1864730_0_-_92	chr5:44649544-44649567(-)	MER121
1866169_0_+_73	chr5:50392064-50392078(+)	MER127
1873731_0_+_53	chr5:58495675-58495729(+)	UCON9
1876134_0_-_86	chr5:60931589-60931609(-)	MER121
1890599_0_+_79	chr5:77309961-77309998(+)	Eulor6B
1895055_0_+_87	chr5:81671505-81671527(+)	MER121
1900625.5_0_+_72	chr5:88585797-88585825(+)	MER121
1902777_0_+_53	chr5:90643387-90643420(+)	AmnSINE1_GG
1903538_0_-_79	chr5:91723955-91723983(-)	MER125
1913159_0_+_89	chr5:103795810-103795856(+)	UCON31
1929440_0_-_73	chr5:124269753-124269778(-)	Eulor5B
1930433_1_+_82	chr5:125337043-125337092(+)	Eulor4
1955119_0_+_44	chr5:145869978-145870018(+)	LF-SINE
1955395_0_+_81	chr5:146094979-146094999(+)	X5A_LINE
1975838_0_-_80	chr5:165688874-165688944(-)	Eulor5A
1979031_0_+_61	chr5:167506770-167506888(+)	Eulor9A
2000476_0_-_85	chr6:8499794-8499914(-)	Eulor6C
2001968_0_-_145	chr6:10178396-10178424(-)	MER131
2008557_0_+_81	chr6:16949073-16949103(+)	UCON26
2031067_0_+_44	chr6:39048083-39048162(+)	Eulor5A
2047829_0_+_41	chr6:55761992-55762018(+)	MER121
2066390_0_-_81	chr6:85075854-85075874(-)	MER121
2078773_0_+_88	chr6:99459607-99459623(+)	UCON9
2080615_0_+_60	chr6:101110199-101110218(+)	MER131
2103549_0_+_93	chr6:128809103-128809145(+)	UCON16
2110337_0_+_44	chr6:136398344-136398375(+)	UCON26
2114503.0_0_+_80	chr6:140364809-140364828(+)	MER121
2115069.5_0_+_82	chr6:141179709-141179763(+)	Eulor5B
2121113_0_+_55	chr6:148347371-148347399(+)	LF-SINE
2122492_0_-_129	chr6:149575272-149575305(-)	SacSINE1
2124929_0_+_75	chr6:152368090-152368109(+)	MER121
2127747_0_+_66	chr6:155834738-155834778(+)	UCON9
2129448_0_-_71	chr6:157252755-157252771(-)	X7B_LINE
2164946_0_-_47	chr7:28364896-28364929(-)	AmnSINE1_GG
2165103_0_+_104	chr7:28447122-28447144(+)	MER121
2177946_0_-_48	chr7:41927983-41928013(-)	MER121
2199161_0_+_76	chr7:78029611-78029627(+)	UCON25
2232211_0_+_45	chr7:113190696-113190791(+)	Eulor6B
2233748_0_-_181	chr7:114618868-114618893(-)	Eulor6B
2265159_0_+_85	chr7:146833245-146833271(+)	UCON4
2289857_0_+_81	chr8:24066738-24066758(+)	MER121
2298526_0_-_61	chr8:33038256-33038278(-)	MER121
2302531_0_+_170	chr8:37341837-37341883(+)	Eulor5B

Table C.2 continued

Name[a]	Coords[b]	TE[c]
2328941_0_+_120	chr8:76879838-76879852(+)	Eulor4
2330918_0_-_108	chr8:79081399-79081462(-)	Eulor3
2339767_0_+_82	chr8:92837069-92837101(+)	MER121
2339841_0_+_84	chr8:92935042-92935090(+)	MER121
2344217_0_+_65	chr8:97188471-97188580(+)	MER135
2354838_0_-_72	chr8:108773965-108774000(-)	UCON26
2379473_0_+_97	chr8:139630131-139630160(+)	Eulor1
2394293_0_+_95	chr9:9338370-9338389(+)	Eulor2A
2397055_0_+_80	chr9:13796837-13796866(+)	Eulor6A
2397055_1_+_73	chr9:13796922-13796943(+)	Eulor6A
2409235_0_+_142	chr9:25589526-25589549(+)	Eulor3
2431229_0_+_103	chr9:75174932-75174961(+)	Eulor6B
2439460_0_-_71	chr9:84169407-84169440(-)	Eulor10
2451401_0_+_100	chr9:101409466-101409483(+)	UCON27
2469220_0_-_74	chr9:118053225-118053243(-)	MER121
2469999_0_+_79	chr9:118715772-118715795(+)	UCON11
2521976_0_-_97	chrX:26365641-26365669(-)	MER131
2551760_0_-_64	chrX:65934376-65934400(-)	MER121
2572113_0_+_91	chrX:98425536-98425558(+)	MER124
2598753_0_+_171	chrX:123865376-123865447(+)	Eulor11
2615332_2_-_77	chrX:137769424-137769462(-)	Eulor5A
236504_0_-_113	chr10:7164642-7164680(-)	MER134
256849_0_+_114	chr10:31429075-31429102(+)	MER131
276892.3_0_-_76	chr10:63322047-63322080(-)	Eulor9A
276891.4_0_+_88	chr10:63322163-63322186(+)	Eulor9A
285555_0_+_63	chr10:72980870-72980944(+)	MER125
292625_0_+_200	chr10:78081072-78081104(+)	Eulor3
292626_0_-_70	chr10:78081151-78081173(-)	Eulor3
295479_0_+_56	chr10:80319880-80319927(+)	MER125
318452_0_-_79	chr10:103288332-103288350(-)	AmnSINE1_GG, AmnSINE1_HS
327752_0_-_82	chr10:111678355-111678382(-)	MER121
330818_0_-_148	chr10:114305414-114305436(-)	MER134
335779_0_+_54	chr10:118027456-118027512(+)	Eulor6D
337561_0_+_69	chr10:119614151-119614186(+)	MER121
338610_0_-_64	chr10:120269645-120269680(-)	Eulor2A
369668_0_-_112	chr11:12893042-12893058(-)	UCON31
373787_0_+_88	chr11:16322974-16322989(+)	MER131
487071_1_+_53	chr11:131453807-131453846(+)	MER121
487071_2_+_103	chr11:131453921-131453949(+)	MER121
512063_0_+_83	chr12:23491193-23491210(+)	LF-SINE
513095_0_+_110	chr12:24372647-24372676(+)	UCON15
516227_0_+_65	chr12:27866363-27866396(+)	AmnSINE1_HS

Table C.2 continued

Name[a]	Coords[b]	TE[c]
551096_0_-_85	chr12:64538090-64538148(-)	Eulor5A
596281_0_+_59	chr12:114732788-114732836(+)	MER121
596947_0_+_93	chr12:115505370-115505426(+)	MER123
622745_0_+_100	chr13:35243065-35243085(+)	MER131
645738_0_-_105	chr13:63899169-63899187(-)	UCON7
645910_0_-_48	chr13:64241844-64241876(-)	MER131
646225_0_+_88	chr13:65068325-65068348(+)	Eulor6B
647102_0_-_48	chr13:66105298-66105320(-)	LF-SINE
648333_0_+_100	chr13:67341863-67341891(+)	Eulor5B
649736_0_-_74	chr13:70683133-70683151(-)	UCON15
651061_0_+_58	chr13:71771123-71771160(+)	Eulor8
651179.1_0_+_60	chr13:71904255-71904296(+)	Eulor6B
653822_0_-_62	chr13:74296760-74296798(-)	Eulor5A
677411_0_+_89	chr13:106457940-106457957(+)	UCON9
677411_1_+_76	chr13:106457967-106457999(+)	UCON9
677411_2_+_85	chr13:106458051-106458083(+)	UCON9
677677_0_+_73	chr13:106991355-106991369(+)	MER123
692404_0_-_138	chr14:28918314-28918337(-)	MER133B
693415_0_+_104	chr14:29848578-29848603(+)	Eulor6A
697653_0_+_69	chr14:33093444-33093479(+)	UCON11
700890_0_-_65	chr14:35855217-35855366(-)	Eulor6A
714186_0_-_107	chr14:53128581-53128608(-)	MER121
751714_0_-_85	chr14:84779620-84779653(-)	UCON30
753005_0_+_40	chr14:86595882-86595921(+)	UCON30
783832_0_-_73	chr15:33865137-33865158(-)	MER133A
785532_1_-_63	chr15:34979215-34979241(-)	MER121
785532_0_-_103	chr15:34979391-34979424(-)	MER121
786420_0_-_57	chr15:35436668-35436697(-)	MER133A
787092_0_-_65	chr15:35993736-35993832(-)	Eulor5A
798275_0_+_115	chr15:43865498-43865524(+)	X7C_LINE
821528_0_-_96	chr15:64357134-64357156(-)	UCON3
853348_0_-_69	chr15:90707731-90707784(-)	UCON7
861597_0_+_94	chr15:98233883-98233900(+)	Eulor1
872174_0_-_78	chr16:5965366-5965388(-)	Eulor3
881780_0_-_82	chr16:17002623-17002644(-)	Eulor6A
905560_0_-_64	chr16:50405915-50405939(-)	MER121
905867_0_+_68	chr16:50695398-50695444(+)	MER121
917018_0_-_100	chr16:60639392-60639413(-)	MER121
920401_0_+_83	chr16:64684847-64684892(+)	Eulor10
931567_0_+_51	chr16:72032498-72032534(+)	MER135
984571_0_+_105	chr17:30189303-30189321(+)	MER123
1015325_0_+_70	chr17:50225698-50225730(+)	Eulor8
1024163_0_+_84	chr17:56744693-56744735(+)	UCON8

Table C.2 continued

Name[a]	Coords[b]	TE[c]
1036020_0_-_69	chr17:66781338-66781363(-)	MER121
1038462_1_-_124	chr17:68911481-68911497(-)	UCON7
1066259_0_+_153	chr18:21498030-21498080(+)	Eulor8
1068544_0_-_89	chr18:23659572-23659590(-)	Eulor2A
1083195_0_+_73	chr18:41146705-41146745(+)	UCON21
1084969_0_+_97	chr18:42888291-42888326(+)	MER123
1091966_0_-_67	chr18:51316122-51316169(-)	MER125
1092513_0_+_129	chr18:51746153-51746176(+)	MER121
1105916_0_-_78	chr18:71369451-71369514(-)	UCON11
1106836_0_-_55	chr18:72345054-72345073(-)	Eulor9C
1107891_0_+_101	chr18:74472718-74472919(+)	MER123
1136689_2_+_91	chr19:35356712-35356733(+)	Eulor5B
1136695_0_+_93	chr19:35387180-35387194(+)	UCON28
1137640_0_-_88	chr19:35962919-35962951(-)	MER121
1137668_1_-_133	chr19:35986024-35986041(-)	UCON18
1137917_0_+_156	chr19:36199466-36199483(+)	Eulor1
1138452_0_-_84	chr19:36589246-36589288(-)	Eulor10
1139164_0_-_65	chr19:37257771-37257787(-)	UCON18
1427009_0_+_75	chr20:37524951-37524982(+)	MER121
1427160_0_-_120	chr20:37678080-37678099(-)	MER125
1427182_0_-_100	chr20:37689430-37689447(-)	MER125
1431929_0_+_129	chr20:41294085-41294112(+)	MER128
1443968_0_-_61	chr20:53838763-53838824(-)	UCON29
1462483_0_+_84	chr21:29777827-29777864(+)	MER127
1463946_0_-_55	chr21:31573085-31573124(-)	Eulor10
1504557_0_+_88	chr22:41047017-41047040(+)	Eulor9A

[a]Name of the EvoFold locus from the hg18 UCSC Genome Browser annotation
[b]Genome coordinates and strand of the EvoFold locus
[c]Name of the co-located TE

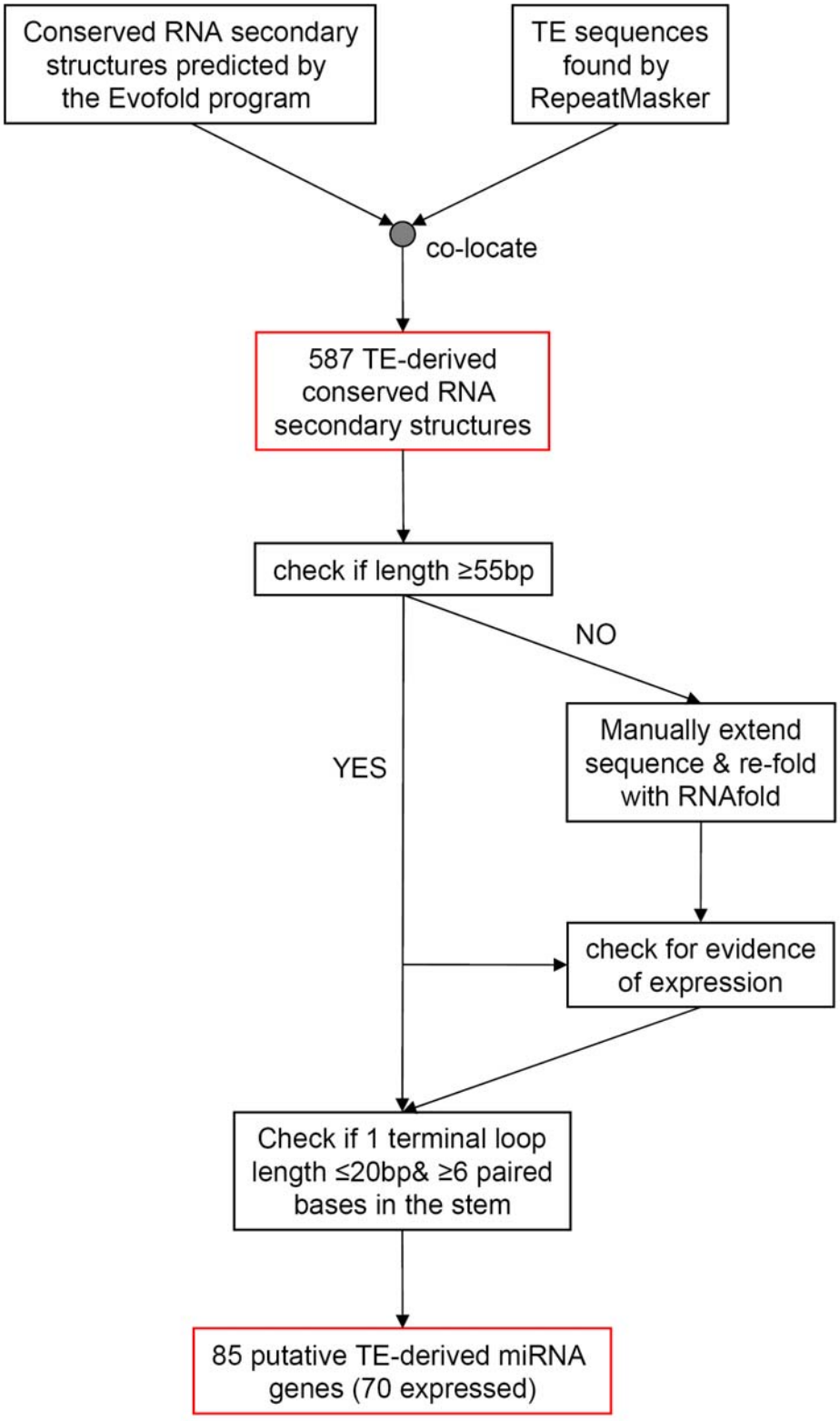

Figure C.1: Protocol for the *ab initio* prediction of human TE-derived miRNA genes

Figure C.2: Genomic structure of TE-derived human miRNAs. Schematics representing the relationships between human miRNA and TE sequences are shown. The first line of each entry summarizes the genomic location of the miRNA along with its association with TEs. The following fields are shown: entry number, miRNA name and accession number, human genome strand, chromosome, start coordinate, end coordinate, miRNA gene length, start and end positions of the mature miRNA sequence, associated TE name and family, the fraction of the mature miRNA sequence, the fraction of TE-derived positions in the miRNA gene, the fraction of TE-derived positions in the mature miRNA sequence. For each entry, a line diagram displays the relationship (overlap) between the miRNA and TE sequences. The miRNA gene, with mature sequence marked as "O", is compared with the locations of TE-derived residues at the same genomic position. The corresponding TE sequences along with their name (family, class) are shown below the composite sequence used for the comparison. The orientation of the TE in the same and opposite strand relative to the miRNA sequences is represented by "X" and "Y", respectively.

```
>>> 4   hsa-mir-562    MI0003568      +       2       232745607      232745701      95      61:80,  L1MB7   L1      20/95(21.05)
        95/95(100.00)  20/20(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOO---------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L1MB7 (L1, LINE)


>>> 5   hsa-mir-566    MI0003572      +       3       50185763       50185856       94      16:34,  AluSg   Alu     19/94(20.21)
        94/94(100.00)  19/19(100.00)

---------------OOOOOOOOOOOOOOOOOO-------------------------------------------------------------------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      AluSg (Alu, SINE)


>>> 6   hsa-mir-28     MI0000086      +       3       189889263      189889348      86      14:35,  L2, L2  L2, L2  22/86(25.58)
        80/86(93.02)   22/22(100.00)

-------------OOOOOOOOOOOOOOOOOOOOO-----------------------------------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----------------------------------------------------      L2 (L2, LINE)
-----------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      L2 (L2, LINE)


>>> 7   hsa-mir-570    MI0003577      +       3       196911452      196911548      97      61:82,  MADE1   Mariner 22/97(22.68)
        80/97(82.47)   22/22(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOO---------------
-----------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX------

-----------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX------      MADE1 (Mariner, DNA)


>>> 8   hsa-mir-571    MI0003578      +       4       333946  334041  96      61:81,  L1MA9, L1MA9   L1, L1  21/96(21.88)   93/96(96.88)
        21/21(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOO---------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYY---YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----------------------------------------------------------      L1MA9 (L1, LINE)
---------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L1MA9 (L1, LINE)
```

Figure C.2 continued

```
>>> 9  hsa-mir-95     MI0000097      -       4       8057928 8058008 81      49:70,  L2, L2  L2, L2  22/81(27.16)   77/81(95.06)
       22/22(100.00)

-----------OOOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY---XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-------------------------------------      L2 (L2, LINE)
----------------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-      L2 (L2, LINE)


>>> 10 hsa-mir-575    MI0003582      -       4       83893514        83893607        94      61:79,  MIR     MIR     19/94(20.21)
       58/94(61.70)   0/19(0.00)

---------------OOOOOOOOOOOOOOOOOO-----------------------------------------------------------------
-----------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

-----------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      MIR (MIR, SINE)


>>> 11 hsa-mir-576    MI0003583      +       4       110629303       110629400       98      16:38,  L1MB7   L1      23/98(23.47)
       98/98(100.00)  23/23(100.00)

---------------OOOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L1MB7 (L1, LINE)


>>> 12 hsa-mir-578    MI0003585      +       4       166526844       166526939       96      61:81,  L2      L2      21/96(21.88)
       43/96(44.79)   21/21(100.00)

-------------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO----------------
-----------------------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

-----------------------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      L2 (L2, LINE)


>>> 13 hsa-mir-579    MI0003586      -       5       32430241        32430338        98      61:83,  MADE1, L1MB8   Mariner, L1
       23/98(23.47)   98/98(100.00)  23/23(100.00)

---------------OOOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY----      MADE1 (Mariner, DNA)
-----------------------------------------------------------------------------------------YYYYY      L1MB8 (L1, LINE)
```

Figure C.2 continued

```
>>> 14  hsa-mir-581    MI0003588      -      5      53283091       53283186       96     16:36,  Charlie10      MER1_type
       21/96(21.88)   4/96(4.17)     0/21(0.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO----------------
XXXX--------------------------------------------------------------------------------------------

XXXX----------------------------------------------------------------------------------------      Charlie10 (MER1_type, DNA)


>>> 15  hsa-mir-582    MI0003589      -      5      59035189       59035286       98     16:38,  L3, L3  CR1, CR1       23/98(23.47)
       84/98(85.71)   23/23(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO----------------
--------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

--------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX------------------------------------------------------      L3 (CR1, LINE)
----------------------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L3 (CR1, LINE)


>>> 16  hsa-mir-584    MI0003591      -      5      148422069      148422165      97     16:37,  MER81   AcHobo  22/97(22.68)
       90/97(92.78)   22/22(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO----------------
-------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

-------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      MER81 (AcHobo, DNA)


>>> 17  hsa-mir-378    MI0000786      +      5      149092581      149092646      66     5:26,44:65,    MIRb, MIRb     MIR, MIR
       44/66(66.67)   60/66(90.91)   43/44(97.73)

----OOOOOOOOOOOOOOOOOOOO-----------------OOOOOOOOOOOOOOOOOOOO-
YYYYYYYYYYYYYYYYYYYYYYY------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

YYYYYYYYYYYYYYYYYYYYYYY----------------------------------------      MIRb (MIR, SINE)
-----------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      MIRb (MIR, SINE)


>>> 18  hsa-mir-585    MI0003592      -      5      168623183      168623276      94     61:79,  MLT1C   MaLR    19/94(20.21)
       29/94(30.85)   14/19(73.68)

---------------OOOOOOOOOOOOOOOOOO-------------------------------------------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYY---------------------------------------------------------------------

YYYYYYYYYYYYYYYYYYYYYYYYYYY-----------------------------------------------------------------------      MLT1C (MaLR, LTR)
```

Figure C.2 continued

```
>>> 19  hsa-mir-340    MI0000802      -       5       179374909      179375003      95      58:80,  MARNA   Mariner 23/95(24.21)
        39/95(41.05)   0/23(0.00)

---------------OOOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------
-------------------------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY--------

-------------------------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY--------      MARNA (Mariner, DNA)


>>> 20  hsa-mir-548a-1 MI0003593      +       6       18679994       18680090       97      61:82,  MADE1   Mariner 22/97(22.68)
        76/97(78.35)   22/22(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOOO---------------
-----------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY----------

-----------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY----------      MADE1 (Mariner, DNA)


>>> 21  hsa-mir-587    MI0003595      +       6       107338693      107338788      96      16:36,  MER115  Tip100  21/96(21.88)
        96/96(100.00)  21/21(100.00)

---------------OOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------
XXXX
```

```
>>> 24  hsa-mir-548a-2 MI0003598      +       6       135601991       135602087       97      61:82,  LTR16A1, MADE1, LTR16A1       ERVL,
Mariner, ERVL  22/97(22.68)   97/97(100.00)  22/22(100.00)

------------------------------------------------------------------OOOOOOOOOOOOOOOOOOOOO----------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYY-------------------------------------------------------------------------------------------      LTR16A1 (ERVL, LTR)
----------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----      MADE1 (Mariner, DNA)
-------------------------------------------------------------------------------------------YYYYYY      LTR16A1 (ERVL, LTR)


>>> 25  hsa-mir-591    MI0003603      -       7       95686910        95687004        95      16:35,  MER5A   MER1_type      20/95(21.05)
       40/95(42.11)   0/20(0.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOO----------------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX------------------------------------------------------------

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX------------------------------------------------------      MER5A (MER1_type, DNA)


>>> 26  hsa-mir-335    MI0000816      +       7       129923188       129923281       94      16:38,  MIRb    MIR     23/94(24.47)
       2/94(2.13)     0/23(0.00)

---------------OOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------
YY-----------------------------------------------------------------------------------------------

YY-----------------------------------------------------------------------------------------------      MIRb (MIR, SINE)


>>> 27  hsa-mir-548a-3 MI0003612      -       8       105565773       105565869       97      61:82,  MLT1G1, MADE1, MLT1G1  MaLR, Mariner,
MaLR   22/97(22.68)   97/97(100.00)  22/22(100.00)

---------------OOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------
YYYYXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXYYYYYYYYYYYYY

YYYYY-------------------------------------------------------------------------------------------      MLT1G1 (MaLR, LTR)
----XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX------------      MADE1 (Mariner, DNA)
-----------------------------------------------------------------------------------YYYYYYYYYYYYY      MLT1G1 (MaLR, LTR)
```

Figure C.2 continued

```
>>> 28  hsa-mir-548d-1 MI0003668      -       8       124429455      124429551      97      61:82,  MADE1   Mariner 22/97(22.68)
        81/97(83.51)   22/22(100.00)

---------------OOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------------
----YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY------------

----YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY------------      MADE1 (Mariner, DNA)


>>> 29  hsa-mir-151    MI0000809      -       8       141811845      141811934      90      46:67,  L2, L2  L2, L2  22/90(24.44)
        90/90(100.00)  22/22(100.00)

-----------------------OOOOOOOOOOOOOOOOOOOOO-----------------------------------------------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX---      L2 (L2, LINE)
-------------------------------------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L2 (L2, LINE)


>>> 30  hsa-mir-421    MI0003685      -       X       73354937       73355021       85      48:70,  L2, L2  L2, L2  23/85(27.06)
        76/85(89.41)   22/23(95.65)

---------------OOOOOOOOOOOOOOOOOOOOO-------------------------------------------------
----YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

----YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-------------------------------------------------      L2 (L2, LINE)
-----------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      L2 (L2, LINE)


>>> 31  hsa-mir-545    MI0003516      -       X       73423664       73423769       106     62:83,  L2, L2  L2, L2  22/106(20.75)
        87/106(82.08)  22/22(100.00)

-----------------------OOOOOOOOOOOOOOOOOOOOO-----------------------------------------------------------------
---------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY----XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

---------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-------------------------------------------------------------      L2 (L2, LINE)
-----------------------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      L2 (L2, LINE)
```

Figure C.2 continued

```
>>> 32  hsa-mir-374    MI0000782      -      X      73423846       73423917       72     12:33,  L2     L2     22/72(30.56)
       39/72(54.17)   0/22(0.00)

----------------------------------------OOOOOOOOOOOOOOOOOOOOO-----------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX--------------------------------

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX--------------------------------      L2 (L2, LINE)


>>> 33  hsa-mir-325    MI0000824      -      X      76142220       76142317       98     16:38,  L2, L2  L2, L2  23/98(23.47)
       88/98(89.80)   23/23(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOOO---------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY----------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY----------------------------------------------------------      L2 (L2, LINE)
----------------------------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      L2 (L2, LINE)


>>> 34  hsa-mir-361    MI0000760      -      X      85045297       85045368       72     6:27,   MER5A  MER1_type      22/72(30.56)
       59/72(81.94)   20/22(90.91)

-----------------------------------------------OOOOOOOOOOOOOOOOOOOOO-----
------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-------

------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-------      MER5A (MER1_type, DNA)


>>> 35  hsa-mir-652    MI0003667      +      X      109185213      109185310      98     61:83,  MER91C  Tip100  23/98(23.47)
       98/98(100.00)  23/23(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOOO---------------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      MER91C (Tip100, DNA)


>>> 36  hsa-mir-513-1  MI0003191      -      X      146102673      146102801      129    37:58,  MER91C  Tip100  22/129(17.05)
       129/129(100.00)        22/22(100.00)

-------------------------------------------------------------------------OOOOOOOOOOOOOOOOOOOOO-----------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
MER91C (Tip100, DNA)
```

Figure C.2 continued

```
>>> 37  hsa-mir-513-2  MI0003192      -      X      146115036      146115162      127     36:57,  MER91C  Tip100  22/127(17.32)
       127/127(100.00)       22/22(100.00)

-------------------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO-----------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
MER91C (Tip100, DNA)


>>> 38  hsa-mir-603    MI0003616      +      10      24604620       24604716       97      61:82,  MADE1   Mariner 22/97(22.68)
       82/97(84.54)   22/22(100.00)

-----------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO---------------
----------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----

----------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----      MADE1 (Mariner, DNA)


>>> 39  hsa-mir-606    MI0003619      +      10      76982222       76982317       96      61:81,  L1MCc   L1      21/96(21.88)
       96/96(100.00)  21/21(100.00)

-----------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO---------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L1MCc (L1, LINE)


>>> 40  hsa-mir-607    MI0003620      -      10      98578416       98578511       96      61:81,  MIR, MIR       MIR, MIR
       21/96(21.88)   96/96(100.00)  21/21(100.00)

---------------OOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------------------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-----------------------------------------      MIR (MIR, SINE)
------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      MIR (MIR, SINE)


>>> 41  hsa-mir-608    MI0003621      +      10      102724732      102724831      100     16:40,  L2      L2      25/100(25.00)
       7/100(7.00)    0/25(0.00)

---------------OOOOOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------YYYYYYY

-----------------------------------------------------------------------------------------------YYYYYYY      L2 (L2, LINE)
```

Figure C.2 continued

```
>>> 42  hsa-mir-612    MI0003625      +       11      64968505       64968604       100     16:40,  MIRb    MIR     25/100(25.00)
        9/100(9.00)    0/25(0.00)

---------------OOOOOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------------
XXXXXXXXX--------------------------------------------------------------------------------------------------------

XXXXXXXXX--------------------------------------------------------------------------------------------      MIRb (MIR, SINE)


>>> 43  hsa-mir-326    MI0000808      -       11      74723784       74723878       95      60:79,  Arthur1 Tip100  20/95(21.05)
        12/95(12.63)   0/20(0.00)

----------------OOOOOOOOOOOOOOOOOOO-----------------------------------------------------------------------
XXXXXXXXXXXX-----------------------------------------------------------------------------------------------

XXXXXXXXXXXX-----------------------------------------------------------------------------------------      Arthur1 (Tip100, DNA)


>>> 44  hsa-mir-616    MI0003629      -       12      56199213       56199309       97      16:37,  L2      L2      22/97(22.68)
        97/97(100.00)  22/22(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO---------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L2 (L2, LINE)


>>> 45  hsa-mir-548c   MI0003630      +       12      63302556       63302652       97      61:82,  MADE1   Mariner 22/97(22.68)
        81/97(83.51)   22/22(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO---------------
-----------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-----

-----------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-----      MADE1 (Mariner, DNA)


>>> 46  hsa-mir-619    MI0003633      -       12      107754813      107754911      99      61:84,  L1MC4, AluSx   L1, Alu 24/99(24.24)
        99/99(100.00)  24/24(100.00)

---------------OOOOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-------------------------------------------------------------      L1MC4 (L1, LINE)
---------------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      AluSx (Alu, SINE)
```

Figure C.2 continued

```
>>> 47  hsa-mir-625    MI0003639      +       14      65007573       65007657       85      15:36,  L1MCa   L1      22/85(25.88)
        85/85(100.00)  22/22(100.00)

--------------OOOOOOOOOOOOOOOOOOOOO-----------------------------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L1MCa (L1, LINE)


>>> 48  hsa-mir-345    MI0000825      +       14      99843949       99844046       98      17:37,  MIR     MIR     21/98(21.43)
        39/98(39.80)   21/21(100.00)

----------------OOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-------------------------------------------------------------------

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      MIR (MIR, SINE)


>>> 49  hsa-mir-493    MI0003132      +       14      100405150      100405238      89      16:37,57:77,   L2      L2      43/89(48.31)
        59/89(66.29)   28/43(65.12)

---------------OOOOOOOOOOOOOOOOOOOOO-------------------OOOOOOOOOOOOOOOOOOOO-------------
------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L2 (L2, LINE)


>>> 50  hsa-mir-370    MI0000778      +       14      100447229      100447303      75      48:68,  MIRm    MIR     21/75(28.00)
        75/75(100.00)  21/21(100.00)

-------------------------------------------------OOOOOOOOOOOOOOOOOOOO-------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      MIRm (MIR, SINE)


>>> 51  hsa-mir-487b   MI0003530      +       14      100582545      100582628      84      51:72,  MIR     MIR     22/84(26.19)
        5/84(5.95)     0/22(0.00)

--------------------------------------------------OOOOOOOOOOOOOOOOOOOO------------
-------------------------------------------------------------------------------------XXXXX

-------------------------------------------------------------------------------------XXXXX      MIR (MIR, SINE)
```

Figure C.2 continued

```
>>> 52  hsa-mir-544    MI0003515      +       14      100584748       100584838       91      55:74,  MER5A1  MER1_type       20/91(21.98)
        91/91(100.00)  20/20(100.00)

----------------------------------------------------------OOOOOOOOOOOOOOOOOOO------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      MER5A1 (MER1_type, DNA)


>>> 53  hsa-mir-626    MI0003640      +       15      39771075        39771168        94      61:79,  L1MB8, L1MCa    L1, L1  19/94(20.21)
        53/94(56.38)   4/19(21.05)

-------------------------------------------------------------------OOOOOOOOOOOOOOOOOO----------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----------------------------------------YYYYYYYYYYYYYYYYYY

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-------------------------------------------------------------------      L1MB8 (L1, LINE)
-----------------------------------------------------------------------------YYYYYYYYYYYYYYYYYY      L1MCa (L1, LINE)


>>> 54  hsa-mir-422a   MI0001444      -       15      61950182        61950271        90      11:32,  MIR3    MIR     22/90(24.44)
        90/90(100.00)  22/22(100.00)

-------------------------------------------------------------OOOOOOOOOOOOOOOOOOOO----------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      MIR3 (MIR, SINE)


>>> 55  hsa-mir-549    MI0003679      -       15      78921374        78921469        96      61:81,  MIRb    MIR     21/96(21.88)
        23/96(23.96)   8/21(38.10)

---------------OOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------
YYYYYYYYYYYYYYYYYYYY-----------------------------------------------------------------------------

YYYYYYYYYYYYYYYYYYYY-------------------------------------------------------------------------      MIRb (MIR, SINE)
```

Figure C.2 continued

```
>>> 56  hsa-mir-633    MI0003648      +       17      58375308       58375405       98      61:83,  MIRb    MIR     23/98(23.47)
        98/98(100.00)  23/23(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOOO----------------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      MIRb (MIR, SINE)


>>> 57  hsa-mir-634    MI0003649      +       17      62213652       62213748       97      61:82,  L1ME3A  L1      22/97(22.68)
        47/97(48.45)   22/22(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOOO----------------
--------------------------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

--------------------------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L1ME3A (L1, LINE)


>>> 58  hsa-mir-548d-2 MI0003671      -       17      62898067       62898163       97      61:82,  MADE1   Mariner 22/97(22.68)
        81/97(83.51)   22/22(100.00)

---------------OOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------------
----YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY------------

----YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY------------      MADE1 (Mariner, DNA)


>>> 59  hsa-mir-637    MI0003652      -       19      3912412 3912510 99      61:84,  L1MC4a  L1      24/99(24.24)   39/99(39.39)
        0/24(0.00)

---------------OOOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------
------------------------------------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

------------------------------------------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L1MC4a (L1, LINE)


>>> 60  hsa-mir-640    MI0003655      +       19      19406872       19406967       96      61:81,  MIRb    MIR     21/96(21.88)
        96/96(100.00)  21/21(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOO----------------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      MIRb (MIR, SINE)
```

Figure C.2 continued

```
>>> 61  hsa-mir-641    MI0003656      -       19      45480290       45480388       99      16:39,  MIR3   MIR     24/99(24.24)
        3/99(3.03)     0/24(0.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOOOOOO----------------
-------------------------------------------------------------------------------------------------YYY

-------------------------------------------------------------------------------------------------YYY      MIR3 (MIR, SINE)


>>> 62  hsa-mir-330    MI0000803      -       19      50834092       50834185       94      57:79,  MIRm   MIR     23/94(24.47)
        50/94(53.19)   0/23(0.00)

---------------OOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------
---------------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

---------------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX      MIRm (MIR, SINE)


>>> 63  hsa-mir-644    MI0003659      +       20      32517791       32517884       94      61:79,  L1MB3  L1      19/94(20.21)
        58/94(61.70)   0/19(0.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOO----------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----------------------------------

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----------------------------------      L1MB3 (L1, LINE)


>>> 64  hsa-mir-645    MI0003660      +       20      48635730       48635823       94      61:79,  MER1B  MER1_type      19/94(20.21)
        59/94(62.77)   0/19(0.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOO----------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----------------------------------

YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY-----------------------------------      MER1B (MER1_type, DNA)


>>> 65  hsa-mir-648    MI0003663      -       22      16843634       16843727       94      16:34,  L2     L2      19/94(20.21)
        93/94(98.94)   19/19(100.00)

------------------------------------------------------------OOOOOOOOOOOOOOOOOO----------------
-YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

-YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      L2 (L2, LINE)
```

Figure C.2 continued

```
>>> 66  hsa-mir-649    MI0003664      -      22     19718465       19718561       97     61:82,  L1M4, MER8, AluSx     L1, MER2_type,
Alu     22/97(22.68)   97/97(100.00)  22/22(100.00)

---------------OOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------------
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

YYYYYYYYYY-------------------------------------------------------------------------------------------      L1M4 (L1, LINE)
---------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY----------      MER8 (MER2_type, DNA)
-----------------------------------------------------------------------------------------YYYYYYYYYY      AluSx (Alu, SINE)


>>> 67  hsa-mir-130b   MI0000748      +      22     20337593       20337674       82     51:72,  MIRm    MIR     22/82(26.83)
        54/82(65.85)   22/22(100.00)

-----------------------------------------------------OOOOOOOOOOOOOOOOOOOOO----------
----------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

----------------------------YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY      MIRm (MIR, SINE)


>>> 68  hsa-mir-659    MI0003683      -      22     36573631       36573727       97     61:82,  Arthur1 Tip100  22/97(22.68)
        45/97(46.39)   22/22(100.00)

---------------OOOOOOOOOOOOOOOOOOOOO-------------------------------------------------------------------------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-----------------------------------------------------

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX-----------------------------------------------------      Arthur1 (Tip100, DNA)
```


Figure C.2 continued

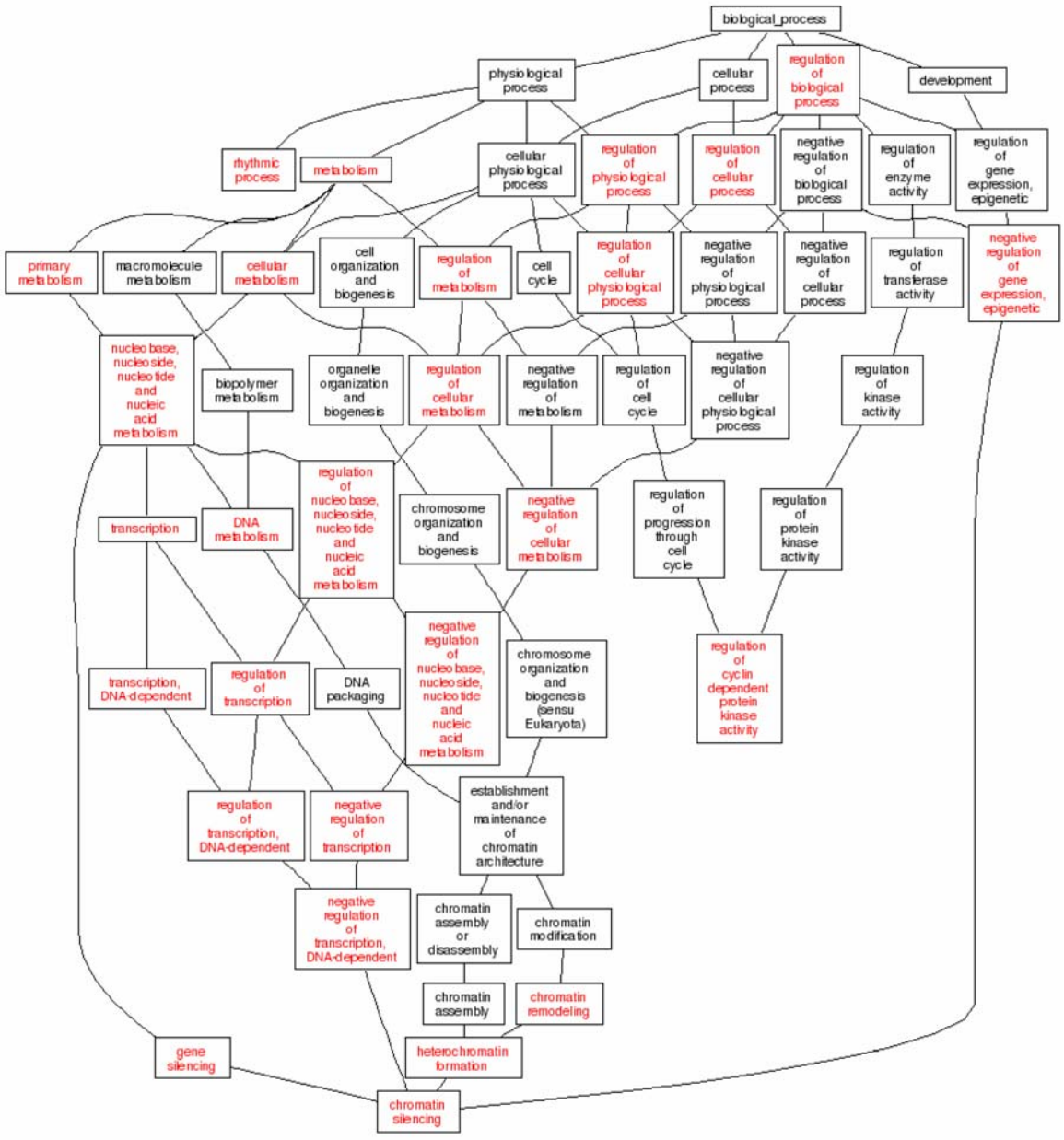

Figure C.3: Gene Ontology (GO) biological process directed acyclic graph showing over-represented GO terms (*P*<0.01; red) associated with mRNA targets of hsa-mir-130b. Targets were identified based on target site complementarity in 3' UTRs and miRNA-mRNA anti-correlated expression patterns.

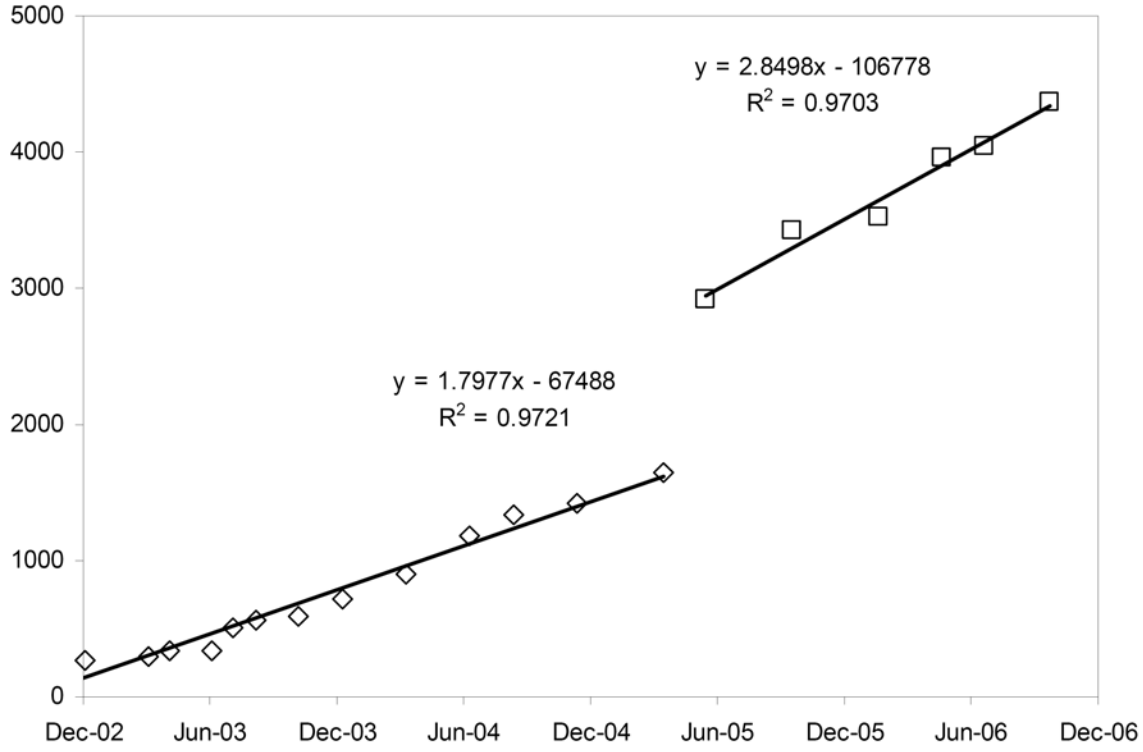

Figure C.4: Rate of increase in the number of miRNA gene entries reported in miRBase. The number of miRNA gene entries is plotted against the database release dates.

APPENDIX D

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Table D.1: Made1 homologous human expressed sequence tags (ESTs)[1]

Hit identifiers	% identity	Length	Mis-matches	Gap openings	Query start	Query end	Hit start	Hit end	E-value	Bit score
gi|23517262|gb|BU674347.1|	94.81	77	4	0	4	80	59	135	7.00E-26	121
gi|11450750|gb|BF438233.1|	93.42	76	5	0	1	76	311	386	7.00E-23	111
gi|18976268|gb|BM668437.1|	93.42	76	5	0	1	76	330	405	7.00E-23	111
gi|19006458|gb|BM693200.1|	93.42	76	5	0	1	76	282	207	7.00E-23	111
gi|19721538|gb|BM996637.1|	93.42	76	5	0	1	76	321	396	7.00E-23	111
gi|23274374|gb|BU608159.1|	93.42	76	5	0	1	76	326	401	7.00E-23	111
gi|2784598|gb|AA743782.1|	93.42	76	5	0	1	76	110	185	7.00E-23	111
gi|2876039|gb|AA804638.1|	93.42	76	5	0	1	76	318	393	7.00E-23	111
gi|3933745|gb|AI290971.1|	93.42	76	5	0	1	76	311	386	7.00E-23	111
gi|4990875|gb|AI702975.1|	93.42	76	5	0	1	76	307	382	7.00E-23	111
gi|5454573|gb|AI832593.1|	93.42	76	5	0	1	76	309	384	7.00E-23	111
gi|7320253|gb|AW615067.1|	93.42	76	5	0	1	76	313	388	7.00E-23	111
gi|8167811|gb|AW976581.1|	93.42	76	5	0	1	76	307	382	7.00E-23	111
gi|8359944|gb|BE042891.1|	93.42	76	5	0	1	76	307	382	7.00E-23	111
gi|52721466|gb|CV371411.1|	95.52	67	3	0	14	80	236	170	3.00E-22	109
gi|32004424|emb|BX492684.1|	92.41	79	6	0	2	80	382	304	3.00E-22	109
gi|2907387|gb|AA833659.1|	93.24	74	5	0	3	76	199	126	1.00E-21	107
gi|52700258|gb|CV350203.1|	93.24	74	5	0	2	75	551	624	1.00E-21	107
gi|6837361|gb|AW340735.1|	93.24	74	5	0	3	76	216	143	1.00E-21	107
gi|7039615|gb|AW469509.1|	93.24	74	5	0	3	76	216	143	1.00E-21	107
gi|3400022|gb|AI073378.1|	91.25	80	7	0	1	80	241	320	2.00E-20	103
gi|46547768|gb|CN478769.1|	91.25	80	7	0	1	80	255	334	2.00E-20	103
gi|20494289|gb|BQ269223.1|	92	75	6	0	1	75	483	409	6.00E-20	101
gi|44842622|gb|CK825697.1|	92	75	6	0	1	75	470	396	6.00E-20	101
gi|45695156|emb|AL519606.3|	94.12	68	3	1	13	80	747	681	3.00E-19	99.6
gi|52811228|gb|CV415725.1|	90.91	77	7	0	4	80	190	266	1.00E-18	97.6
gi|2908283|gb|AA834684.1|	94.12	68	3	1	5	72	137	203	4.00E-18	95.6
gi|13292606|gb|BG399158.1|	90.79	76	7	0	5	80	232	157	4.00E-18	95.6
gi|52653216|gb|CV330002.1|	90.79	76	7	0	5	80	163	88	4.00E-18	95.6
gi|8061011|gb|AW896806.1|	90.79	76	7	0	1	76	289	214	4.00E-18	95.6
gi|52667308|gb|CV344094.1|	92.11	76	5	1	1	76	280	206	4.00E-18	95.6
gi|5438416|gb|AI819337.1|	90	80	8	0	1	80	241	320	4.00E-18	95.6
gi|7946376|gb|AW850859.1|	90	80	8	0	1	80	170	249	4.00E-18	95.6
gi|12766146|gb|BG256330.1|	92.5	80	4	2	1	80	385	308	4.00E-18	95.6
gi|27846682|emb|BX105680.1|	91.55	71	6	0	8	78	364	434	2.00E-17	93.7
gi|3837536|gb|AI242139.1|	91.55	71	6	0	8	78	269	199	2.00E-17	93.7
gi|58568449|dbj|BP395858.1|	90.54	74	7	0	7	80	284	211	6.00E-17	91.7

Table D.1 continued

Hit identifiers	% identity	Length	Mis-matches	Gap openings	Query start	Query end	Hit start	Hit end	E-value	Bit score
gi|6602709|emb|AL134522.1|	93.24	74	3	2	4	77	28	99	6.00E-17	91.7
gi|91749404|gb|EB386059.1|	91.3	69	6	0	11	79	162	94	2.00E-16	89.7
gi|32005544|emb|BX493226.1|	92.75	69	4	1	7	75	216	149	2.00E-16	89.7
gi|14321058|gb|BG926535.1|	90.41	73	7	0	2	74	661	589	2.00E-16	89.7
gi|52707894|gb|CV357839.1|	90.41	73	7	0	2	74	127	55	2.00E-16	89.7
gi|82333517|dbj|DA902558.1|	90.91	77	6	1	1	77	193	118	2.00E-16	89.7
gi|12120877|gb|BF772977.1|	90.12	81	7	1	1	80	235	155	2.00E-16	89.7
gi|12120883|gb|BF772983.1|	90.12	81	7	1	1	80	236	156	2.00E-16	89.7
gi|5110886|gb|AI742598.1|	90.12	81	7	1	1	80	241	321	2.00E-16	89.7
gi|1885842|gb|AA250882.1|	92.19	64	5	0	1	64	41	104	1.00E-15	87.7
gi|82341158|dbj|DB016887.1|	92.19	64	5	0	1	64	188	251	1.00E-15	87.7
gi|10200151|gb|BE778953.1|	90.28	72	7	0	9	80	149	220	1.00E-15	87.7
gi|13339103|gb|BG432597.1|	89.47	76	8	0	5	80	516	441	1.00E-15	87.7
gi|13343062|gb|BG436556.1|	89.47	76	8	0	1	76	300	375	1.00E-15	87.7
gi|15164200|emb|AL600694.1|	89.47	76	8	0	1	76	324	399	1.00E-15	87.7
gi|18983536|gb|BM673638.1|	89.47	76	8	0	5	80	116	41	1.00E-15	87.7
gi|19005651|gb|BM692393.1|	89.47	76	8	0	5	80	185	260	1.00E-15	87.7
gi|2834284|gb|AA774950.1|	89.47	76	8	0	1	76	221	146	1.00E-15	87.7
gi|11977833|gb|BF692425.1|	90.79	76	6	1	1	76	386	460	1.00E-15	87.7
gi|13452873|gb|BG491361.1|	90.79	76	6	1	1	76	8	82	1.00E-15	87.7
gi|13580923|gb|BG573270.1|	90.79	76	6	1	1	76	290	364	1.00E-15	87.7
gi|19727271|gb|BQ002371.1|	90.79	76	6	1	1	76	400	326	1.00E-15	87.7
gi|24776874|gb|CA414223.1|	90.79	76	6	1	1	76	400	326	1.00E-15	87.7
gi|27932373|gb|CB106566.1|	90.79	76	6	1	1	76	13	87	1.00E-15	87.7
gi|28365225|gb|CB243581.1|	90.79	76	6	1	1	76	29	103	1.00E-15	87.7
gi|43429246|emb|BX952415.1|	90.79	76	6	1	5	80	89	163	1.00E-15	87.7
gi|43425548|emb|BX951140.1|	88.75	80	9	0	1	80	136	57	1.00E-15	87.7
gi|3038959|gb|AA903836.1|	90	80	7	1	1	80	74	152	1.00E-15	87.7
gi|5543963|gb|AI869995.1|	90	80	7	1	1	80	450	372	1.00E-15	87.7
gi|7668921|gb|AW753989.1|	91.04	67	6	0	12	78	421	487	4.00E-15	85.7
gi|7668972|gb|AW754040.1|	91.04	67	6	0	12	78	421	487	4.00E-15	85.7
gi|8046501|gb|AW884489.1|	88.61	79	9	0	2	80	132	210	4.00E-15	85.7
gi|14466558|gb|BI059028.1|	90	70	7	0	3	72	129	198	1.00E-14	83.8
gi|81125345|dbj|DA460339.1|	90	70	7	0	3	72	349	280	1.00E-14	83.8
gi|27845181|emb|BX102210.1|	89.74	78	7	1	1	77	407	330	1.00E-14	83.8
gi|31915369|emb|BX479525.1|	89.04	73	8	0	4	76	130	202	6.00E-14	81.8
gi|66791763|dbj|BP425510.1|	89.04	73	8	0	4	76	184	256	6.00E-14	81.8
gi|685935|gb|T71414.1|	88.16	76	9	0	1	76	11	86	6.00E-14	81.8
gi|711241|gb|T82953.1|	88.16	76	9	0	1	76	11	86	6.00E-14	81.8
gi|1404173|gb|W88623.1|	87.5	80	10	0	1	80	146	67	6.00E-14	81.8
gi|1891141|gb|AA257012.1|	88.89	81	8	1	1	80	248	168	6.00E-14	81.8
gi|81181343|dbj|DA639796.1|	90.62	64	6	0	1	64	16	79	2.00E-13	79.8
gi|8058080|gb|AW893875.1|	92.19	64	4	1	1	64	471	409	2.00E-13	79.8
gi|14372680|gb|BG954509.1|	90.28	72	6	1	9	80	238	168	2.00E-13	79.8
gi|79163886|dbj|DA105807.1|	90.28	72	6	1	9	80	387	317	2.00E-13	79.8
gi|8623066|gb|BE160345.1|	90.28	72	6	1	9	80	94	164	2.00E-13	79.8

Table D.1 continued

Hit identifiers	% identity	Length	Mis-matches	Gap openings	Query start	Query end	Hit start	Hit end	E-value	Bit score
gi|8623148|gb|BE160427.1|	90.28	72	6	1	9	80	94	164	2.00E-13	79.8
gi|2162267|gb|AA448597.1|	88.16	76	9	0	1	76	343	418	2.00E-13	79.8
gi|80799866|dbj|DA505931.1|	88.16	76	9	0	1	76	121	196	2.00E-13	79.8
gi|21855046|gb|BQ716149.1|	89.47	76	7	1	1	76	119	193	2.00E-13	79.8
gi|24805094|gb|CA440674.1|	89.47	76	7	1	1	76	400	326	2.00E-13	79.8
gi|83480277|dbj|DB358036.1|	89.47	76	7	1	1	76	382	308	2.00E-13	79.8
gi|3056341|gb|AA916949.1|	87.5	80	10	0	1	80	230	309	2.00E-13	79.8
gi|1764951|gb|AA181484.1|	88.75	80	8	1	2	80	360	281	2.00E-13	79.8
gi|8054117|gb|AW889912.1|	88.75	80	8	1	1	80	149	227	2.00E-13	79.8
gi|83532058|dbj|DB333866.1|	88.75	80	8	1	1	80	65	143	2.00E-13	79.8
gi|10107714|gb|BE719449.1|	88.73	71	8	0	1	71	623	553	9.00E-13	77.8
gi|504666|dbj|D20846.1|	90.14	71	6	1	1	70	191	121	9.00E-13	77.8
gi|14393270|gb|BG989200.1|	89.33	75	7	1	6	80	304	231	9.00E-13	77.8
gi|31446439|gb|CD514721.1|	89.33	75	7	1	6	80	13	86	9.00E-13	77.8
gi|2617003|gb|AA663012.1|	88.46	78	7	1	5	80	105	28	9.00E-13	77.8
gi|90847359|dbj|DB577513.1|	87.65	81	7	1	1	78	52	132	9.00E-13	77.8
gi|83241952|dbj|DB315742.1|	90	70	6	1	11	79	73	142	4.00E-12	75.8
gi|91749668|gb|EB386323.1|	90	70	6	1	9	77	226	157	4.00E-12	75.8
gi|10918992|dbj|AV761144.1|	88.46	78	8	1	3	79	236	313	4.00E-12	75.8
gi|78737823|dbj|DA326471.1|	88.46	78	8	1	3	80	82	6	4.00E-12	75.8
gi|83199537|dbj|DB235269.1|	85.88	85	6	1	2	80	480	396	4.00E-12	75.8
gi|7668920|gb|AW753988.1|	89.23	65	7	0	14	78	91	27	1.00E-11	73.8
gi|7668971|gb|AW754039.1|	89.23	65	7	0	14	78	91	27	1.00E-11	73.8
gi|83486421|dbj|DB358889.1|	89.23	65	7	0	16	80	329	393	1.00E-11	73.8
gi|2328991|gb|AA558514.1|	87.67	73	9	0	4	76	115	43	1.00E-11	73.8
gi|81156387|dbj|DA383600.1|	87.67	73	9	0	4	76	516	588	1.00E-11	73.8
gi|8165082|gb|AW973998.1|	87.67	73	9	0	4	76	244	172	1.00E-11	73.8
gi|23373989|gb|BU661807.1|	88.16	76	7	1	5	80	85	158	1.00E-11	73.8
gi|3896467|gb|AI274199.1|	88.16	76	7	1	1	74	74	149	1.00E-11	73.8
gi|5054918|gb|AI733805.1|	88.16	76	7	1	1	74	72	147	1.00E-11	73.8
gi|82136433|dbj|DB047679.1|	88.31	77	8	1	1	76	152	228	1.00E-11	73.8
gi|33252132|gb|CF136688.1|	90.62	64	5	1	1	64	30	92	6.00E-11	71.9
gi|46922787|emb|BX405577.2|	87.5	80	9	1	1	80	146	224	6.00E-11	71.9
gi|46233530|emb|AL566894.3|	86.3	73	9	1	1	73	526	455	2.00E-10	69.9
gi|3872647|gb|AI264444.1|	86.84	76	8	1	1	74	72	147	2.00E-10	69.9
gi|33258518|gb|CF143074.1|	86.25	80	7	1	1	80	184	259	2.00E-10	69.9
gi|7111293|gb|AW499536.1|	86.25	80	7	1	1	80	184	259	2.00E-10	69.9
gi|7111295|gb|AW499537.1|	86.25	80	7	1	1	80	184	259	2.00E-10	69.9
gi|7111327|gb|AW499553.1|	86.25	80	7	1	1	80	184	259	2.00E-10	69.9
gi|7116331|gb|AW502136.1|	86.25	80	7	1	1	80	184	259	2.00E-10	69.9
gi|7116335|gb|AW502138.1|	86.25	80	7	1	1	80	184	259	2.00E-10	69.9
gi|3214298|gb|AI004788.1|	85.71	84	8	1	1	80	128	45	2.00E-10	69.9
gi|90648194|dbj|BY797461.2|	85.71	84	8	1	1	80	347	430	2.00E-10	69.9
gi|1486755|gb|AA022674.1|	84.88	86	7	1	1	80	283	198	2.00E-10	69.9
gi|1486863|gb|AA022709.1|	84.88	86	7	1	1	80	101	186	2.00E-10	69.9
gi|83124689|dbj|DB343577.1|	86.49	74	10	0	4	77	405	478	9.00E-10	67.9

Table D.1 continued

Hit identifiers	% identity	Length	Mis-matches	Gap openings	Query start	Query end	Hit start	Hit end	E-value	Bit score
gi|90938918|dbj|DB507251.1|	87.84	74	8	1	1	74	74	2	9.00E-10	67.9
gi|694186|gb|T76983.1|	86.84	76	7	1	5	80	204	132	9.00E-10	67.9
gi|83078449|dbj|DB106337.1|	85.9	78	7	1	1	78	458	385	9.00E-10	67.9
gi|83237080|dbj|DB354909.1|	86.25	80	8	1	1	80	218	294	9.00E-10	67.9
gi|12189868|gb|BF837652.1|	88.41	69	7	1	1	69	148	215	4.00E-09	65.9
gi|8167508|gb|AW976282.1|	87.67	73	8	1	5	76	495	423	4.00E-09	65.9
gi|14399447|gb|BG995377.1|	85.71	77	11	0	4	80	227	151	4.00E-09	65.9
gi|81108769|dbj|DA381665.1|	85.9	78	7	1	1	78	487	414	4.00E-09	65.9
gi|83190413|dbj|DB352537.1|	87.65	81	8	2	1	80	312	391	4.00E-09	65.9
gi|83517488|dbj|DB143470.1|	86.84	76	9	1	1	76	116	42	1.00E-08	63.9
gi|82338292|dbj|DB049879.1|	85	80	12	0	1	80	166	87	1.00E-08	63.9
gi|80933029|dbj|DA523524.1|	86.25	80	10	1	1	80	468	390	1.00E-08	63.9

[1]BLASTN was used to search the human EST database with a full length Made1 element query sequence. Only hits that were ≥80% identical over ≥80% of the length of the element are reported. Hit identifiers (Genbank identification numbers and accessions) are shown followed by the BLAST statistics for each query-hit pair.

Table D.2: Over-represented GO biological process categories among genes with Made1-derived hsa-mir-548 target sites

GO ID[a]	Description[b]	Gene acc[c]	Obs[d]	Exp[e]	*P*-value[f]
GO:0000087	M phase of mitotic cell cycle	ENSG00000130177 ENSG00000086827* ENSG00000004897*	3	0.44	9.42E-03
GO:0007067	mitosis	ENSG00000130177 ENSG00000086827* ENSG00000004897*	3	0.43	9.06E-03
GO:0007088	regulation of mitosis	ENSG00000130177 ENSG00000086827*	2	0.12	6.39E-03
GO:0006917	induction of apoptosis	ENSG00000163161 ENSG00000171132* ENSG00000004468	3	0.44	9.42E-03
GO:0012502	induction of programmed cell death	ENSG00000163161 ENSG00000171132* ENSG00000004468	3	0.44	9.42E-03
GO:0008283	cell proliferation	ENSG00000076716 ENSG00000112038* ENSG00000143125 ENSG00000125657 ENSG00000130177 ENSG00000004897*	6	1.58	4.47E-03
GO:0007059	chromosome segregation	ENSG00000163535 ENSG00000086827*	2	0.11	5.17E-03

[a]GO biological process category ID
[b]Functional description for the GO category
[c]The list of Ensembl gene accessions in the GO category, * indicates genes that are down-regulated in colorectal cancer tissue
[d]Observed gene number in the GO category
[e]Expected gene number in the GO category
[f]*P*-value showing significance of enrichment for the GO category based on the hypergeometric test

Table D.3: Over-represented GO biological process categories among genes with miRanda predicted hsa-mir-548 target sites that map to colorectal cancer down-regulated co-expression clusters (*i.e.* 12, 15 & 20 in Figure 5.6).

GO ID[a]	Description[b]	Gene acc[c]	Obs[d]	Exp[e]	*P*-value[f]
GO:0007155	cell adhesion	ENSG00000179776 ENSG00000040731 ENSG00000154162 ENSG00000133800 ENSG00000073712 ENSG00000138080 ENSG00000018236 ENSG00000038427 ENSG00000170989 ENSG00000146648 ENSG00000128536 ENSG00000087303 ENSG00000115414 ENSG00000102290 ENSG00000164171 ENSG00000158887 ENSG00000067141 ENSG00000124215 ENSG00000107562 ENSG00000112378 ENSG00000143341 ENSG00000164199 ENSG00000077522 ENSG00000104415 ENSG00000163347 ENSG00000154655 ENSG00000198542	27	10.05	2.61E-06
GO:0016337	cell-cell adhesion	ENSG00000179776 ENSG00000040731 ENSG00000154162 ENSG00000146648 ENSG00000128536	10	3.45	2.48E-03

Table D.3 continued

GO ID[a]	Description[b]	Gene acc[c]	Obs[d]	Exp[e]	*P*-value[f]
		ENSG00000102290 ENSG00000158887 ENSG00000124215 ENSG00000164199 ENSG00000163347			
GO:0007156	homophilic cell adhesion	ENSG00000179776 ENSG00000040731 ENSG00000154162 ENSG00000128536 ENSG00000102290 ENSG00000158887 ENSG00000124215	7	2.16	5.99E-03
GO:0031589	cell-substrate adhesion	ENSG00000133800 ENSG00000087303 ENSG00000164171 ENSG00000077522 ENSG00000198542	5	0.9	2.03E-03
GO:0007160	cell-matrix adhesion	ENSG00000133800 ENSG00000087303 ENSG00000164171 ENSG00000077522 ENSG00000198542	5	0.9	2.03E-03
GO:0007154	cell communication	ENSG00000064989 ENSG00000153208 ENSG00000145632 ENSG00000166073 ENSG00000184984 ENSG00000080644 ENSG00000147432 ENSG00000135902 ENSG00000108018 ENSG00000018236 ENSG00000174429 ENSG00000169676 ENSG00000170989 ENSG00000146648 ENSG00000140009 ENSG00000151348 ENSG00000138685 ENSG00000115641 ENSG00000115414 ENSG00000113327 ENSG00000091844 ENSG00000164949 ENSG00000146072 ENSG00000135821 ENSG00000127920 ENSG00000177464	76	51.71	1.01E-04

Table D.3 continued

GO ID[a]	Description[b]	Gene acc[c]	Obs[d]	Exp[e]	*P*-value[f]
		ENSG00000132975			
		ENSG00000064652			
		ENSG00000171189			
		ENSG00000095752			
		ENSG00000164171			
		ENSG00000183111			
		ENSG00000182634			
		ENSG00000113594			
		ENSG00000101665			
		ENSG00000116141			
		ENSG00000124089			
		ENSG00000143198			
		ENSG00000158887			
		ENSG00000067141			
		ENSG00000134259			
		ENSG00000170485			
		ENSG00000133636			
		ENSG00000165588			
		ENSG00000169860			
		ENSG00000167941			
		ENSG00000115252			
		ENSG00000154678			
		ENSG00000172572			
		ENSG00000113448			
		ENSG00000108551			
		ENSG00000156475			
		ENSG00000156218			
		ENSG00000144724			
		ENSG00000115665			
		ENSG00000166592			
		ENSG00000107562			
		ENSG00000196632			
		ENSG00000196781			
		ENSG00000105989			
		ENSG00000175868			
		ENSG00000152284			
		ENSG00000182880			
		ENSG00000164199			
		ENSG00000104415			
		ENSG00000049246			
		ENSG00000124104			
		ENSG00000078043			
		ENSG00000165970			
		ENSG00000149305			
		ENSG00000170579			
		ENSG00000198542			
		ENSG00000137962			
		ENSG00000198752			
		ENSG00000064692			

Table D.3 continued

GO ID[a]	Description[b]	Gene acc[c]	Obs[d]	Exp[e]	*P*-value[f]
		ENSG00000198929			
GO:0007267	cell-cell signaling	ENSG00000153208 ENSG00000166073 ENSG00000147432 ENSG00000169676 ENSG00000140009 ENSG00000138685 ENSG00000135821 ENSG00000171189 ENSG00000095752 ENSG00000158887 ENSG00000067141 ENSG00000134259 ENSG00000115665 ENSG00000107562 ENSG00000104415 ENSG00000165970 ENSG00000149305 ENSG00000170579 ENSG00000064692 ENSG00000198929	20	8.45	3.13E-04
GO:0019226	transmission of nerve impulse	ENSG00000166073 ENSG00000147432 ENSG00000169676 ENSG00000135821 ENSG00000171189 ENSG00000158887 ENSG00000115665 ENSG00000165970 ENSG00000149305 ENSG00000170579 ENSG00000064692 ENSG00000198929	12	3.79	4.30E-04
GO:0007268	synaptic transmission	ENSG00000166073 ENSG00000147432 ENSG00000169676 ENSG00000135821 ENSG00000171189 ENSG00000158887 ENSG00000115665 ENSG00000165970 ENSG00000149305 ENSG00000170579 ENSG00000064692 ENSG00000198929	12	3.64	3.02E-04
GO:0001505	regulation of neurotransmitter levels	ENSG00000135821 ENSG00000115665 ENSG00000064692	4	0.73	5.84E-03

Table D.3 continued

GO ID[a]	Description[b]	Gene acc[c]	Obs[d]	Exp[e]	*P*-value[f]
		ENSG00000198929			
GO:0007165	signal transduction	ENSG00000064989 ENSG00000153208 ENSG00000145632 ENSG00000166073 ENSG00000184984 ENSG00000080644 ENSG00000147432 ENSG00000135902 ENSG00000108018 ENSG00000018236 ENSG00000174429 ENSG00000169676 ENSG00000170989 ENSG00000146648 ENSG00000140009 ENSG00000151348 ENSG00000138685 ENSG00000115641 ENSG00000115414 ENSG00000113327 ENSG00000091844 ENSG00000164949 ENSG00000146072 ENSG00000127920 ENSG00000177464 ENSG00000132975 ENSG00000064652 ENSG00000171189 ENSG00000164171 ENSG00000183111 ENSG00000182634 ENSG00000113594 ENSG00000101665 ENSG00000116141 ENSG00000124089 ENSG00000143198 ENSG00000170485 ENSG00000133636 ENSG00000165588 ENSG00000169860 ENSG00000167941 ENSG00000115252 ENSG00000154678 ENSG00000172572 ENSG00000113448 ENSG00000108551 ENSG00000156475 ENSG00000156218	65	47.47	2.90E-03

Table D.3 continued

GO ID[a]	Description[b]	Gene acc[c]	Obs[d]	Exp[e]	*P*-value[f]
		ENSG00000144724 ENSG00000166592 ENSG00000107562 ENSG00000196632 ENSG00000196781 ENSG00000105989 ENSG00000175868 ENSG00000152284 ENSG00000182880 ENSG00000164199 ENSG00000104415 ENSG00000049246 ENSG00000124104 ENSG00000078043 ENSG00000198542 ENSG00000137962 ENSG00000198752			
GO:0051056	regulation of small GTPase mediated signal transduction	ENSG00000174429 ENSG00000183111 ENSG00000198752	3	0.27	2.38E-03
GO:0035023	regulation of Rho protein signal transduction	ENSG00000174429 ENSG00000183111	2	0.06	1.52E-03
GO:0007266	Rho protein signal transduction	ENSG00000174429 ENSG00000183111 ENSG00000137962	3	0.39	6.51E-03
GO:0009966	regulation of signal transduction	ENSG00000145632 ENSG00000174429 ENSG00000091844 ENSG00000183111 ENSG00000165588 ENSG00000167941 ENSG00000196781 ENSG00000152284 ENSG00000198752	9	3.31	6.04E-03
GO:0006575	amino acid derivative metabolism	ENSG00000129596 ENSG00000131480 ENSG00000115665 ENSG00000064692	4	0.82	9.08E-03
GO:0009250	glucan biosynthesis	ENSG00000111713 ENSG00000056998	2	0.15	8.65E-03
GO:0005978	glycogen biosynthesis	ENSG00000111713 ENSG00000056998	2	0.15	8.65E-03
GO:0007417	central nervous system development	ENSG00000061676 ENSG00000171189 ENSG00000170485 ENSG00000165588 ENSG00000134595	6	1.82	9.91E-03

Table D.3 continued

GO ID[a]	Description[b]	Gene acc[c]	Obs[d]	Exp[e]	*P*-value[f]
		ENSG00000043355			
GO:0007596	blood coagulation	ENSG00000095752 ENSG00000164171 ENSG00000169860 ENSG00000143341 ENSG00000154655	5	1.29	9.39E-03
GO:0051260	protein homooligomerization	ENSG00000187134 ENSG00000077522	2	0.15	8.65E-03
GO:0050952	sensory perception of electrical stimulus	ENSG00000182634 ENSG00000182880	2	0	0.00E+00
GO:0050978	magnetoreception, using electrical stimulus	ENSG00000182634 ENSG00000182880	2	0	0.00E+00
GO:0050954	sensory perception of mechanical stimulus	ENSG00000153208 ENSG00000115380 ENSG00000131480 ENSG00000140522 ENSG00000143341 ENSG00000164199	6	1.58	5.03E-03
GO:0050979	magnetoreception, using mechanical stimulus	ENSG00000153208 ENSG00000115380 ENSG00000131480 ENSG00000140522 ENSG00000143341 ENSG00000164199	6	0	0.00E+00
GO:0019233	sensory perception of pain	ENSG00000165091 ENSG00000164199 ENSG00000136156	3	0.03	1.41E-06
GO:0050966	detection of mechanical stimulus during sensory perception of pain	ENSG00000165091 ENSG00000164199 ENSG00000136156	3	0	0.00E+00
GO:0051341	regulation of oxidoreductase activity	ENSG00000146648 ENSG00000198929	2	0.06	1.52E-03
GO:0050999	regulation of nitric-oxide synthase activity	ENSG00000146648 ENSG00000198929	2	0.06	1.52E-03

[a]GO biological process category ID
[b]Functional description for the GO category
[c]The list of Ensembl gene accessions in the GO category
[d]Observed gene number in the GO category
[e]Expected gene number in the GO category
[f]*P*-value showing significance of enrichment for the GO category based on the hypergeometric test

Table D.4: Putative hsa-mir-548 target genes previously implicated as being involved in colorectal cancer by microarray expression profiling

Accn[a]	Ref[b]	Name[c]	Status[d]	Target[e]	*P*-value[f]
ENST00000282050	(TAKEMASA *et al.* 2001)	ATP synthase alpha chain, mitochondrial precursor (EC 3.6.3.14)	down	a	1.73E-05
ENST00000219660	(TAKEMASA *et al.* 2001)	Aquaporin-8 (AQP-8)	down	b	0.0050
ENST00000262825	(KITAHARA *et al.* 2001)	Cytokine receptor common beta chain precursor (GM-CSF/IL-3/IL-5 receptor common beta-chain) (CD131 antigen) (CDw131)	down	b	0.0006
ENST00000201031	(KITAHARA *et al.* 2001)	Transcription factor AP-2 gamma (AP2-gamma) (Activating enhancer- binding protein 2 gamma) (Transcription factor ERF-1)	down	a,b,c,d	0.0006
ENST00000241261	(KITAHARA *et al.* 2001)	Tumor necrosis factor ligand superfamily member 10 (TNF-related apoptosis-inducing ligand) (TRAIL protein) (Apo-2 ligand) (Apo-2L) (CD253 antigen)	down	a	0.0315
ENST00000360121	(KITAHARA *et al.* 2001)	Leukosialin precursor (Leucocyte sialoglycoprotein) (Sialophorin) (Galactoglycoprotein) (GALGP) (CD43 antigen)	down	a,c	0.0018
ENST00000360876	(KITAHARA *et al.* 2001)	Eukaryotic translation initiation factor 3 subunit 9 (eIF-3 eta) (eIF3 p116) (eIF3 p110) (eIF3b) (Prt1 homolog) (hPrt1)	up	a	0.0176
ENST00000368083	(BERTUCCI *et al.* 2004)	Arginase-1 (EC 3.5.3.1) (Type I arginase) (Liver-type arginase)	down	c	0.0308
ENST00000344548	(BERTUCCI *et al.* 2004)	Cell division control protein 42 homolog precursor (G25K GTP-binding protein)	down	c	0.0455
ENST00000379328	(BERTUCCI *et al.* 2004)	Trans-acting T-cell-specific transcription factor GATA-3 (GATA-binding factor 3)	down	a	0.0012
ENST00000285900	(BERTUCCI *et al.* 2004)	Glutamate receptor 1 precursor (GluR-1) (GluR-A) (GluR-K1) (Glutamate receptor ionotropic, AMPA 1) (AMPA-selective glutamate receptor 1)	down	d	0.0058

Table D.4 continued

Accn[a]	Ref[b]	Name[c]	Status[d]	Target[e]	*P*-value[f]
ENST00000328245	(BERTUCCI *et al.* 2004)	Heat shock factor protein 1 (HSF 1) (Heat shock transcription factor 1) (HSTF 1)	down	c	1.03E-05
ENST00000227752	(BERTUCCI *et al.* 2004)	Interleukin-10 receptor alpha chain precursor (IL-10R-A) (IL-10R1) (CDw210a antigen)	down	d	0.0206
ENST00000371794	(BERTUCCI *et al.* 2004)	Noelin precursor (Neuronal olfactomedin-related ER localized protein) (Olfactomedin-1)	down	d	0.0410
ENST00000334661	(BERTUCCI *et al.* 2004)	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta 1 (EC 3.1.4.11) (Phosphoinositide phospholipase C) (PLC-delta-1) (Phospholipase C-delta-1) (PLC-III)	down	c	0.0198
ENST00000229390	(BERTUCCI *et al.* 2004)	Splicing factor, arginine/serine-rich 9 (Pre-mRNA-splicing factor SRp30C)	down	a	0.0005
ENST00000340600	(BERTUCCI *et al.* 2004)	Suppressor of cytokine signaling 2 (SOCS-2) (Cytokine-inducible SH2 protein 2) (CIS-2) (STAT-induced STAT inhibitor 2) (SSI-2)	down	b	0.0056
ENST00000288207	(BERTUCCI *et al.* 2004)	G2/mitotic-specific cyclin-B2	up	a,b,c,d	0.0010
ENST00000264161	(BERTUCCI *et al.* 2004)	Aspartyl-tRNA synthetase (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS)	up	c	0.0121
ENST00000309268	(BERTUCCI *et al.* 2004)	Elongation factor 1-alpha 1 (EF-1-alpha-1) (Elongation factor 1 A-1) (eEF1A-1) (Elongation factor Tu) (EF-Tu)	up	a	0.0252
ENST00000319974	(BERTUCCI *et al.* 2004)	no description (ets variant gene 4 (E1A enhancer binding protein, E1AF))	up	a	0.0025
ENST00000302068	(BERTUCCI *et al.* 2004)	Fibrinogen beta chain precursor [Contains: Fibrinopeptide B]	up	a,c	0.0053
ENST00000341048	(BERTUCCI *et al.* 2004)	no description (interleukin 6 signal transducer (gp130, oncostatin M receptor))	up	a	0.0004

Table D.4 continued

Accn[a]	Ref[b]	Name[c]	Status[d]	Target[e]	*P*-value[f]
ENST00000296585	(BERTUCCI *et al.* 2004)	Integrin alpha-2 precursor (Platelet membrane glycoprotein Ia) (GPIa) (Collagen receptor) (VLA-2 alpha chain) (CD49b antigen)	up	d	0.0088
ENST00000260302	(BERTUCCI *et al.* 2004)	Collagenase 3 precursor (EC 3.4.24.-) (Matrix metalloproteinase-13) (MMP-13)	up	b	0.0197
ENST00000296930	(BERTUCCI *et al.* 2004)	Nucleophosmin (NPM) (Nucleolar phosphoprotein B23) (Numatrin) (Nucleolar protein NO38)	up	a	0.0015
ENST00000216392	(BERTUCCI *et al.* 2004)	Glycogen phosphorylase, liver form (EC 2.4.1.1)	up	b,c,d	0.0467
ENST00000370321	(BERTUCCI *et al.* 2004)	60S ribosomal protein L5	up	c	0.0032
ENST00000265361	(BERTUCCI *et al.* 2004)	Semaphorin-3C precursor (Semaphorin E) (Sema E)	up	d	8.05E-05
ENST00000244520	(BERTUCCI *et al.* 2004)	U1 small nuclear ribonucleoprotein C (U1 snRNP protein C) (U1C protein) (U1-C)	up	a	0.0001
ENST00000273258	(KWON *et al.* 2004)	PRA1 family protein 3 (ARL-6-interacting protein 5) (ADP-ribosylation- like factor 6-interacting protein 5) (Aip-5) (Glutamate transporter EAAC1-interacting protein) (GTRAP3-18) (Prenylated Rab acceptor protein 2) (Protein JWa) (Dermal papilla-derived protein 11)	down	b	0.0014
ENST00000323456	(KWON *et al.* 2004)	myotubularin related protein 4	down	b	0.0011
ENST00000258428	(KWON *et al.* 2004)	DNA repair protein REV1 (EC 2.7.7.-) (Rev1-like terminal deoxycytidyl transferase) (Alpha integrin-binding protein 80) (AIBP80)	down	c	0.0011
ENST00000326361	(KWON *et al.* 2004)	Zinc finger protein 639 (Zinc finger protein ZASC1) (Zinc finger protein ANC_2H01)	up	a	0.0015
ENST00000259075	(KWON *et al.* 2004)	TRAF family member-associated NF-kappa-B activator (TRAF-interacting protein) (I-TRAF)	up	b,c,d	7.5E-05

Table D.4 continued

Accn[a]	Ref[b]	Name[c]	Status[d]	Target[e]	*P*-value[f]
ENST00000262462	(KWON *et al.* 2004)	Long-chain fatty acid transport protein 6 (Fatty acid transport protein 6) (FATP-6) (Very long-chain acyl-CoA synthetase homolog 1) (VLCSH1) (hVLCS-H1) (Fatty-acid-coenzyme A ligase, very long-chain 2) (Solute carrier family 27 member 6)	up	a,c	6.4E-05
ENST00000307633	(KWON *et al.* 2004)	Histidyl-tRNA synthetase (EC 6.1.1.21) (Histidine--tRNA ligase) (HisRS)	up	c	0.0001
ENST00000327304	(KWON *et al.* 2004)	Exosome complex exonuclease RRP40 (EC 3.1.13.-) (Ribosomal RNA- processing protein 40) (Exosome component 3) (p10)	up	b,c,d	0.0001
ENST00000370986	(KWON *et al.* 2004)	Growth arrest and DNA-damage-inducible protein GADD45 alpha (DNA-damage-inducible transcript 1) (DDIT1)	up	a	0.0013
ENST00000160827	(KWON *et al.* 2004)	Kinesin-like protein KIF22 (Kinesin-like DNA-binding protein) (Kinesin-like protein 4)	up	a	0.0031
ENST00000230588	(NOTTERMAN *et al.* 2001)	Meprin A subunit alpha precursor (EC 3.4.24.18) (Endopeptidase-2) (N- benzoyl-L-tyrosyl-P-amino-benzoic acid hydrolase subunit alpha) (PABA peptide hydrolase) (PPH alpha)	down	a	0.0012
ENST00000162749	(NOTTERMAN *et al.* 2001)	Tumor necrosis factor receptor superfamily member 1A precursor (p60) (TNF-R1) (TNF-RI) (TNFR-I) (p55) (CD120a antigen) [Contains: Tumor necrosis factor receptor superfamily member 1A, membrane form; Tumor necrosis factor-binding protein 1 (TBPI)]	down	b	0.0023
ENST00000314355	(NOTTERMAN *et al.* 2001)	Cyclin-dependent kinases regulatory subunit 2 (CKS-2)	up	a	0.0136
ENST00000283646	(SHIH *et al.* 2005)	Ribose-5-phosphate isomerase (EC 5.3.1.6) (Phosphoriboisomerase)	down	a	4.2E-05

Table D.4 continued

Accn[a]	Ref[b]	Name[c]	Status[d]	Target[e]	*P*-value[f]
ENST00000356245	(SHIH *et al.* 2005)	Ras-GTPase-activating protein-binding protein 1 (EC 3.6.1.-) (ATP- dependent DNA helicase VIII) (GAP SH3-domain-binding protein 1) (G3BP- 1) (HDH-VIII)	up	b	4.15E-05

[a]Ensembl transcript accession for putative hsa-mir-548 target genes
[b]Publication where the genes involvement in colorectal cancer was originally reported
[c]Name and brief description of the gene
[d]Expression status of the gene (up- or down-regulated) in colorectal cancer relative to normal tissue
[e]Paralog-specific hsa-mir-548 target site
[f]*P*-value associated with the hsa-mir-548 target sites

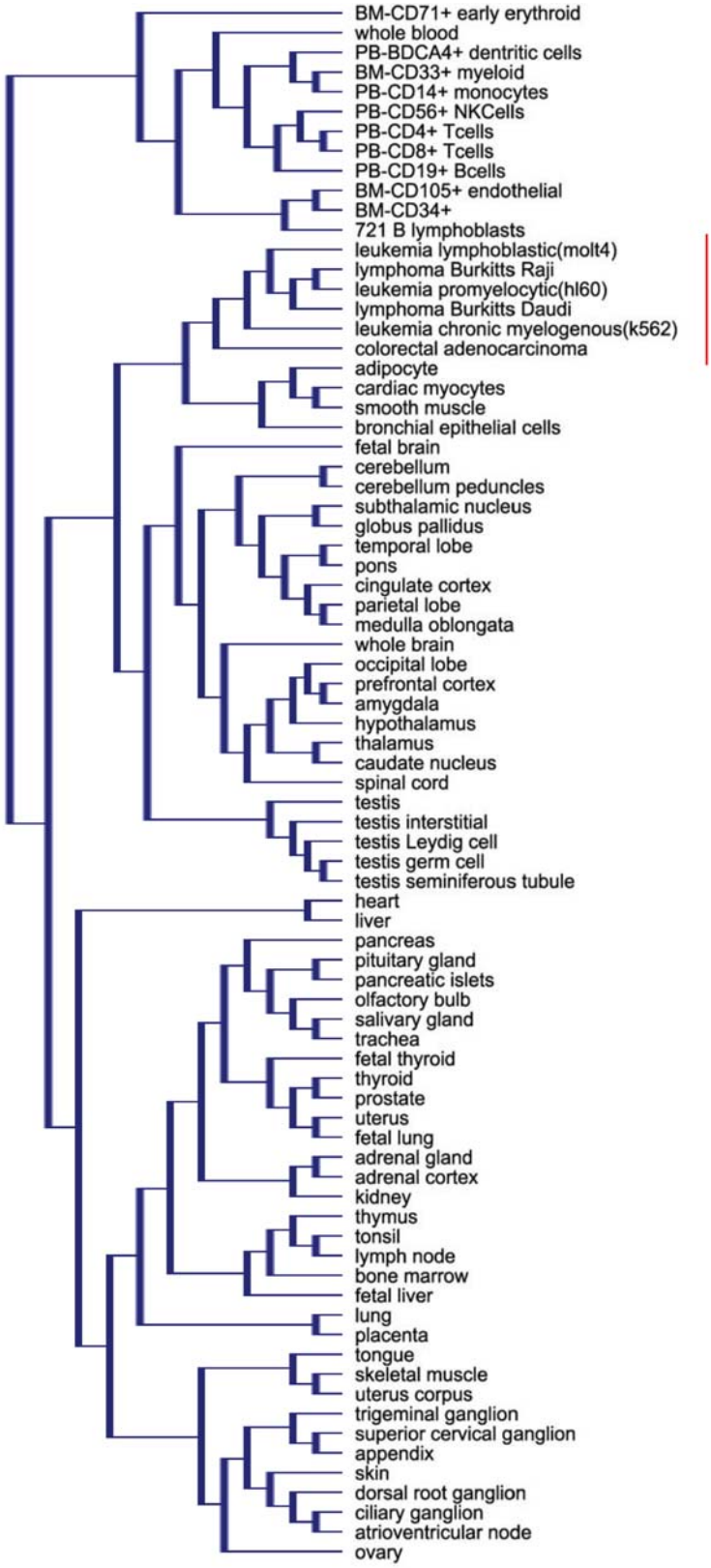

Figure D.1: Dendogram showing relationships among tissues from the Novartis Foundation Symatlas microarray dataset. Cancer tissues are indicated with the red bar.

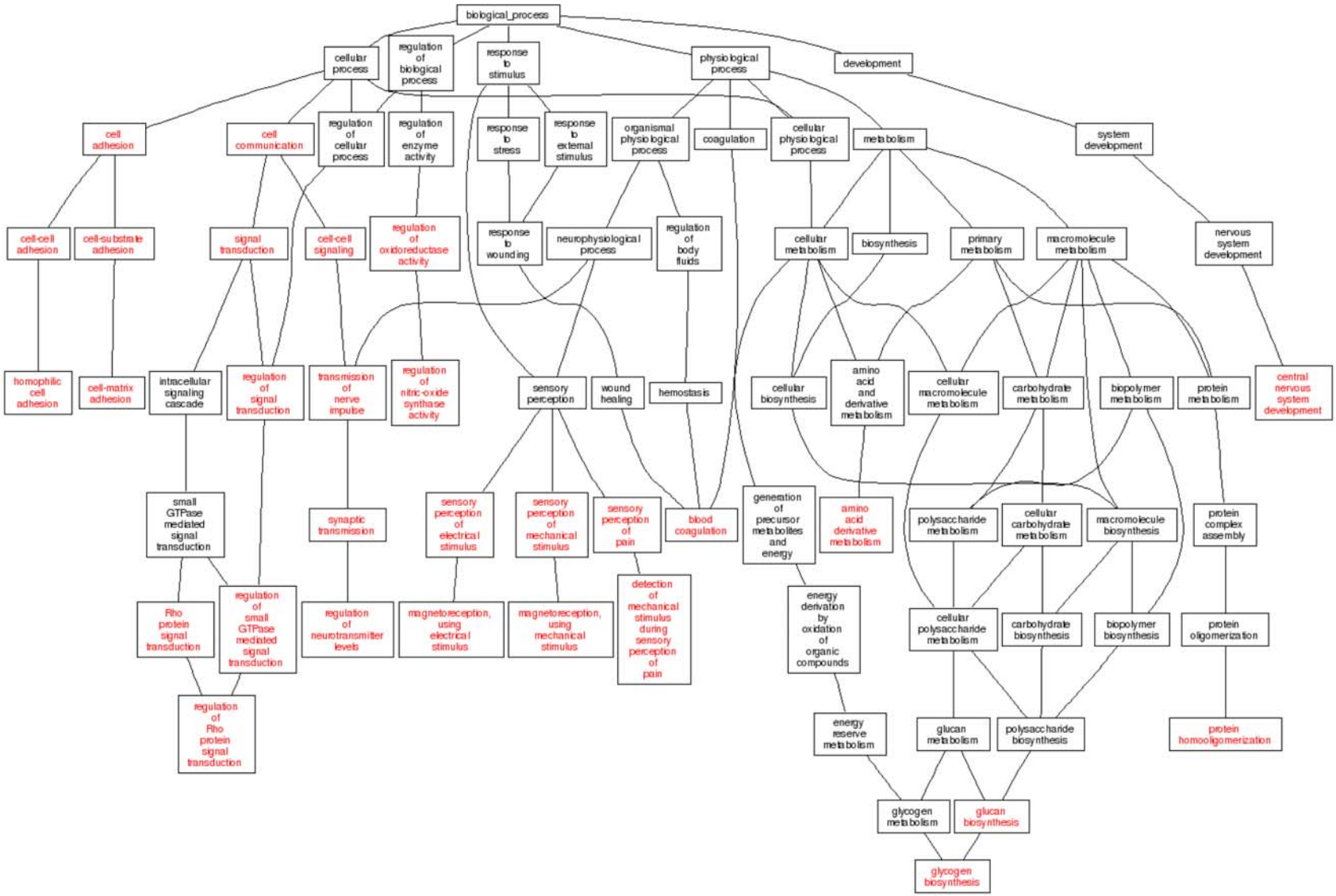
biological_process
cellular process
regulation of biological process
response to stimulus
physiological process
development
cell adhesion
cell communication
regulation of cellular process
regulation of enzyme activity
response to stress
response to external stimulus
organismal physiological process
coagulation
cellular physiological process
metabolism
system development
cell-cell adhesion
cell-substrate adhesion
signal transduction
cell-cell signaling
regulation of oxidoreductase activity
response to wounding
neurophysiological process
regulation of body fluids
cellular metabolism
biosynthesis
primary metabolism
macromolecule metabolism
nervous system development
homophilic cell adhesion
cell-matrix adhesion
intracellular signaling cascade
regulation of signal transduction
transmission of nerve impulse
regulation of nitric-oxide synthase activity
sensory perception
wound healing
hemostasis
cellular biosynthesis
amino acid and derivative metabolism
cellular macromolecule metabolism
carbohydrate metabolism
biopolymer metabolism
protein metabolism
central nervous system development
small GTPase mediated signal transduction
synaptic transmission
sensory perception of electrical stimulus
sensory perception of mechanical stimulus
sensory perception of pain
blood coagulation
generation of precursor metabolites and energy
amino acid derivative metabolism
polysaccharide metabolism
cellular carbohydrate metabolism
macromolecule biosynthesis
protein complex assembly
Rho protein signal transduction
regulation of small GTPase mediated signal transduction
regulation of neurotransmitter levels
magnetoreception, using electrical stimulus
magnetoreception, using mechanical stimulus
detection of mechanical stimulus during sensory perception of pain
energy derivation by oxidation of organic compounds
cellular polysaccharide metabolism
carbohydrate biosynthesis
biopolymer biosynthesis
protein oligomerization
regulation of Rho protein signal transduction
energy reserve metabolism
glucan metabolism
polysaccharide biosynthesis
protein homooligomerization
glycogen metabolism
glucan biosynthesis
glycogen biosynthesis

Figure D.2: Over-represented GO biological process categories among genes with miRanda predicted hsa-mir-548 target sites that map to colorectal cancer down-regulated co-expression clusters (*i.e.* 12, 15 & 20 in Figure 5.6). The portion of the directed acyclic graph (DAG) containing all paths from the root biological process term to the over-represented functional category terms is shown. Over-represented functional categories are indicated in red.

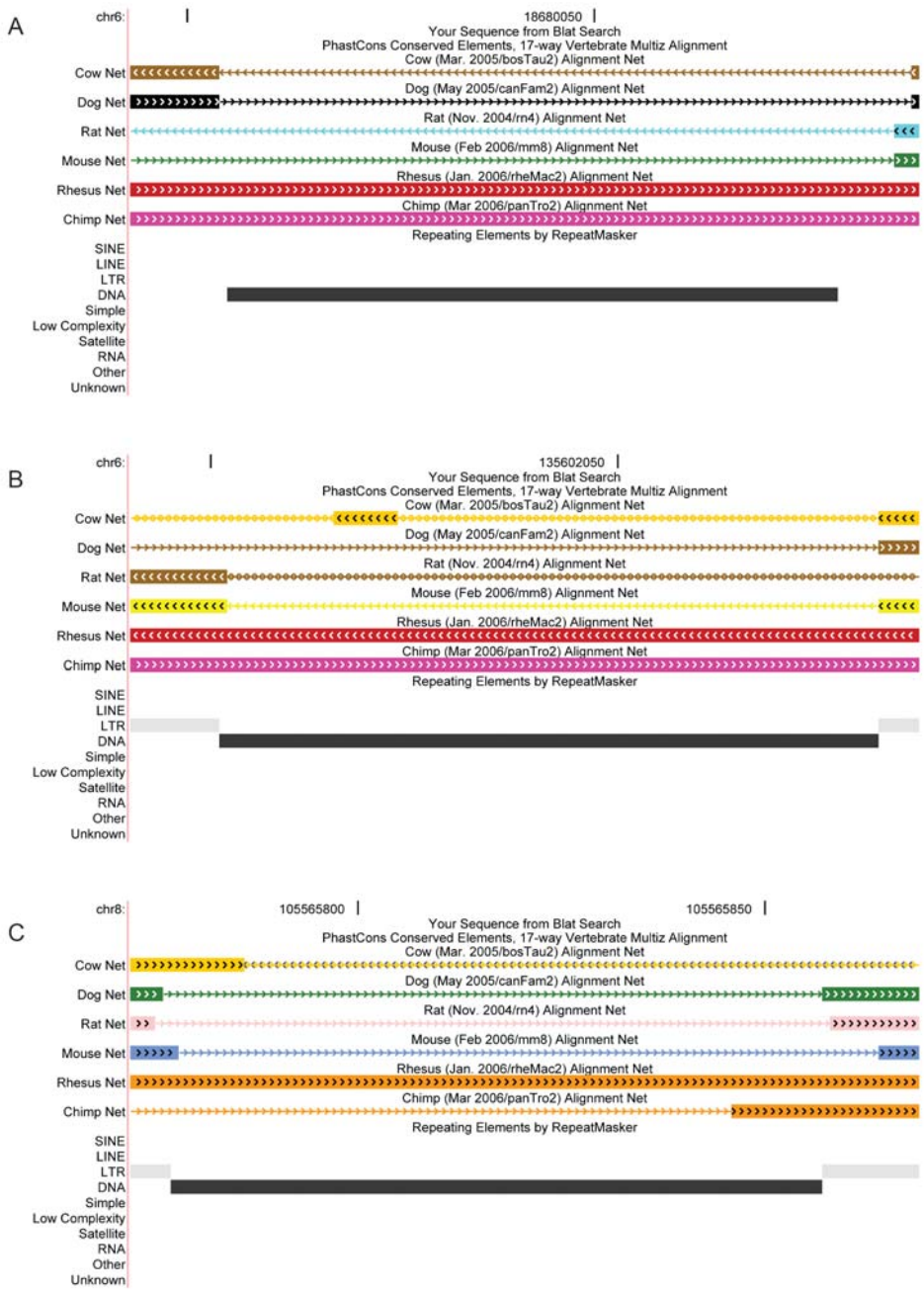

Figure D.3: Made1-derived miRNA genes are primate-specific. Human genomic regions corresponding to Made1-derived miRNA genes are shown: (A) hsa-mir-548a-1, (B) hsa-mir-548a-2, (C) hsa-mir-548a-3, (D) hsa-mir-548b, (E) hsa-mir-548c, (F) hsa-mir-548d-1, (G) hsa-mir-548d-2. The UCSC Genome Browser is used to show the location of the Made1 elements (DNA) in the RepeatMasker track. Evolutionary comparisons between the human genome and the corresponding regions in the chimp, rhesus, mouse, rat, dog and cow genomes are shown using the species-specific Net tracks of the Genome Browser. Corresponding Made1 orthologous regions that are present in another species are indicated with a broad line, while regions that are missing in another species are shown with a thin line.

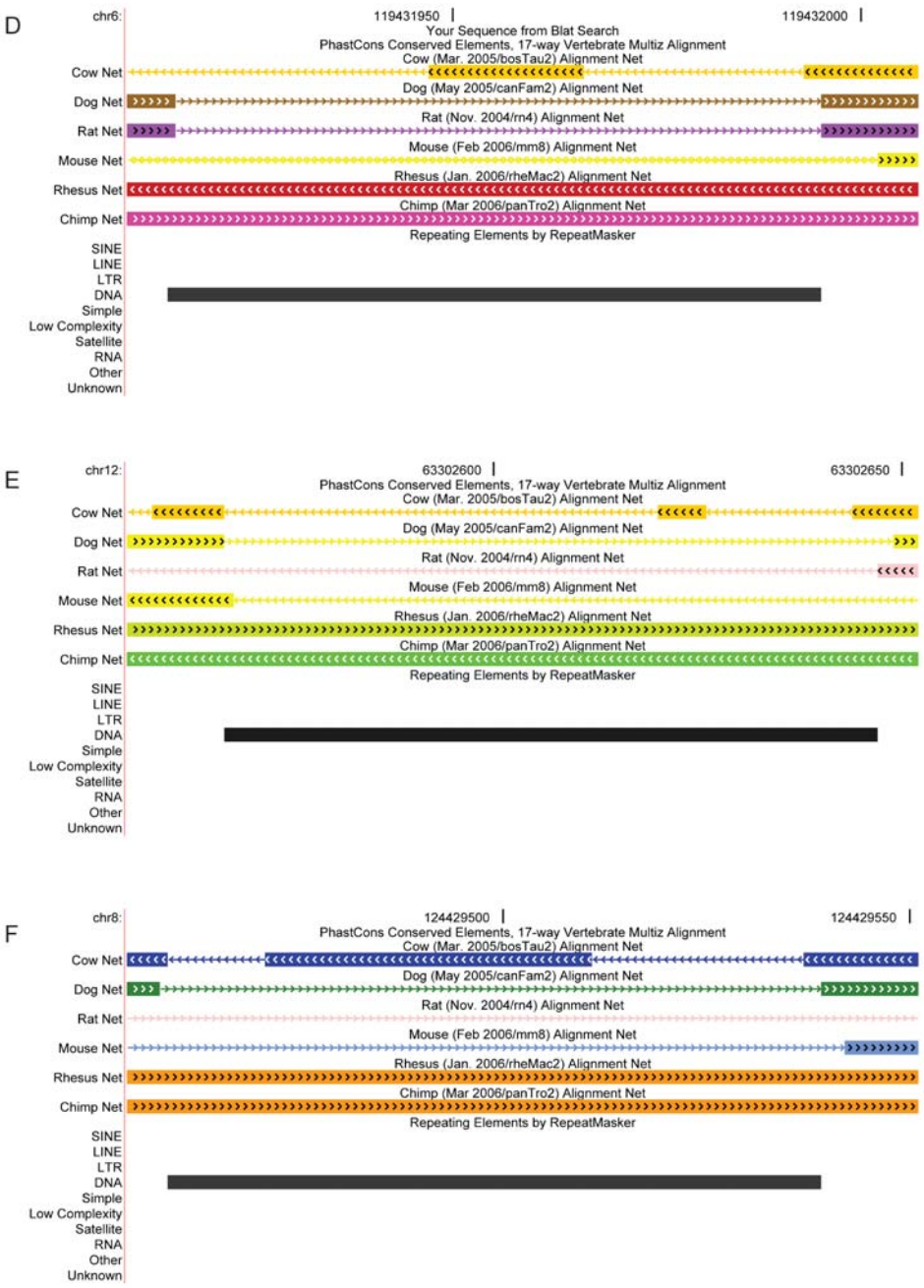

Figure D.3 continued

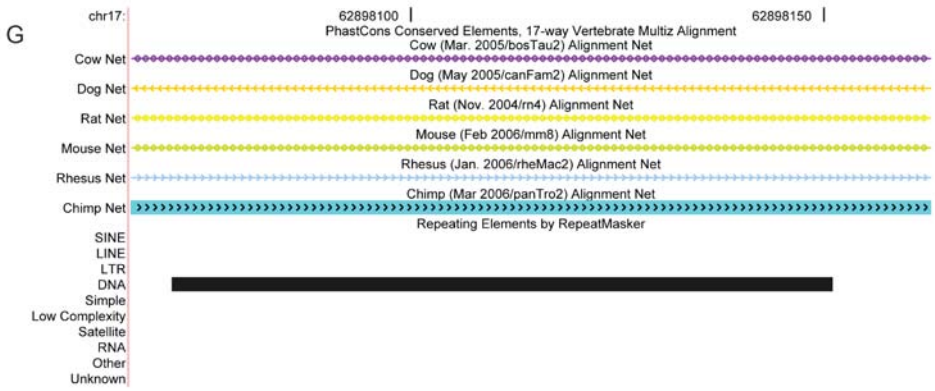

Figure D.3 continued

APPENDIX E

SUPPLEMENTARY INFORMATION FOR CHAPTER 6

Table E.1: TE-derived miRNAs

Table E.1 continued

Name[a]	Accn[b]	Coords[c]	TE[d]	Overlap[e]
ath-MIR414	MI0001425	chr1:25141119-25141226(-)	ATCopia24I (LTR/Copia)	57.41
ath-MIR855	MI0005411	chr2:4681509-4681780(+)	Athila4B_LTR (LTR/Gypsy)	100.00
ath-MIR416	MI0001427	chr2:7015602-7015681(+)	Vandal1 (DNA/MuDR)	100.00
ath-MIR405a	MI0001074	chr2:9642037-9642193(-)	SIMPLEHAT2 (DNA/hAT)	100.00
ath-MIR407	MI0001079	chr2:13873282-13873544(+)	ATMU9 (DNA/MuDR)	93.16
ath-MIR405d	MI0001077	chr4:2789653-2789738(-)	SIMPLEHAT2 (DNA/hAT)	100.00
ath-MIR401	MI0001070	chr4:5020234-5020483(-)	Athila4B_LTR (LTR/Gypsy)	100.00
ath-MIR854b	MI0005413	chr5:11341600-11341820(-)	Athila6A_I (LTR/Gypsy)	100.00
ath-MIR854d	MI0005415	chr5:11707091-11707311(-)	Athila6A_I (LTR/Gypsy)	100.00
ath-MIR854c	MI0005414	chr5:11855326-11855546(+)	Athila6A_I (LTR/Gypsy)	100.00
ath-MIR854a	MI0005412	chr5:11864949-11865169(+)	Athila6A_I (LTR/Gypsy)	100.00
ath-MIR405b	MI0001075	chr5:20649740-20649863 (+)	SIMPLEHAT2 (DNA/hAT)	100.00
osa-MIR439a	MI0001691	chr1:20206990-20207082(+)	MuDR4_OS (DNA/MuDR)	100.00
osa-MIR814a	MI0005239	chr1:22701877-22701973(+)	STOWAWAY47_OS (DNA/Stowaway)	100.00
osa-MIR441c	MI0001706	chr1:32522492-32522645(+)	STOWAWAY1_OS (DNA/Stowaway)	97.40
osa-MIR812a	MI0005233	chr1:34273999-34274232(+)	STOWAWAY51_OS (DNA/Stowaway)	100.00
osa-MIR807a	MI0005209	chr1:39312844-39313095(+)	ECR (DNA/Tourist)	98.41
osa-MIR818a	MI0005247	chr1:40742271-40742414(+)	STOWAWAY15-2_OS (DNA/Stowaway)	98.61
osa-MIR819a	MI0005252	chr1:41534243-41534367(+)	STOWAWAY1_OS (DNA/Stowaway)	100.00
osa-MIR806a	MI0005210	chr1:43254846-43255097(-)	TREP215 (DNA/Stowaway)	94.44

Table E.1 continued

Name[a]	Accn[b]	Coords[c]	TE[d]	Overlap[e]
osa-MIR812b	MI0005234	chr2:1936324-1936493(-)	STOWAWAY51_OS (DNA/Stowaway)	100.00
osa-MIR818b	MI0005248	chr2:4007187-4007299(+)	STOWAWAY15-2_OS (DNA/Stowaway)	100.00
osa-MIR806b	MI0005211	chr2:5044109-5044323(-)	TREP215 (DNA/Stowaway)	100.00
osa-MIR814c	MI0005241	chr2:10889670-10889752(-)	STOWAWAY47_OS (DNA/Stowaway)	100.00
osa-MIR817	MI0005246	chr2:12276361-12276443(-)	ENSPM3_OS (DNA/En-Spm)	100.00
osa-MIR437	MI0001688	chr2:17044466-17044678(-)	STOWAWAY41_OS (DNA/Stowaway)	37.09
osa-MIR807b	MI0005218	chr2:24481931-24482076(-)	ECR (DNA/Tourist)	100.00
osa-MIR814b	MI0005240	chr2:26335342-26335415(+)	STOWAWAY47_OS (DNA/Stowaway)	100.00
osa-MIR819c	MI0005254	chr2:33750674-33750827(+)	STOWAWAY1_OS (DNA/Stowaway)	97.40
osa-MIR819b	MI0005253	chr2:34659352-34659517(+)	DITAILA (DNA/Tourist), STOWAWAY50_OS (DNA/Stowaway)	96.39
osa-MIR818c	MI0005249	chr2:35922869-35923044(+)	STOWAWAY46_OS (DNA/Stowaway)	89.77
osa-MIR808	MI0005220	chr3:8847036-8847187(+)	STOWAWAY1_OS (DNA/Stowaway)	98.68
osa-MIR819d	MI0005255	chr3:10848548-10848699(-)	STOWAWAY1_OS (DNA/Stowaway), STOWAWAY10_OS (DNA/Stowaway)	100.00
osa-MIR439d	MI0001694	chr3:13677143-13677240(+)	MuDR4_OS (DNA/MuDR)	89.80
osa-MIR435	MI0001687	chr3:18164352-18164483(+)	MERMITEH (DNA)	39.39
osa-MIR821a	MI0005266	chr3:22928833-22929106(+)	ENSPM3_OS (DNA/En-Spm), OSTE22 (DNA)	100.00
osa-MIR809a	MI0005221	chr3:26735515-26735675(-)	STOWAWAY1_OS (DNA/Stowaway)	94.41
osa-MIR441a	MI0001704	chr3:28876745-28876897(-)	STOWAWAY1_OS (DNA/Stowaway)	79.74
osa-MIR443	MI0001708	chr3:29972009-29972156(+)	STOWAWAY47_OS (DNA/Stowaway)	100.00
osa-MIR806c	MI0005212	chr3:36133235-36133504(-)	TREP215 (DNA/Stowaway)	88.52
osa-MIR819e	MI0005256	chr3:36206839-36206992(-)	STOWAWAY1_OS (DNA/Stowaway)	97.40
osa-MIR420	MI0001440	chr4:6098543-6098697(+)	TRUNCATOR2_OS (LTR/Gypsy)	100.00

Table E.1 continued

Name[a]	Accn[b]	Coords[c]	TE[d]	Overlap[e]
osa-MIR416	MI0001436	chr4:17268776-17268884(+)	CPSC3_LTR (LTR/Copia)	100.00
osa-MIR807c	MI0005219	chr4:23886344-23886527(+)	ECR (DNA/Tourist)	100.00
osa-MIR442	MI0001707	chr4:32149607-32149839(+)	OLO24B (DNA/Tourist)	100.00
osa-MIR806d	MI0005213	chr4:33568288-33568558(-)	TREP215 (DNA/Stowaway)	87.82
osa-MIR818d	MI0005250	chr4:34750094-34750232(-)	TREP220 (DNA/Stowaway)	81.29
osa-MIR819f	MI0005257	chr4:35070636-35070779(-)	STOWAWAY50_OS (DNA/Stowaway)	100.00
osa-MIR815b	MI0005243	chr5:14914142-14914271(+)	WANDERER_OS (DNA/Tourist)	54.62
osa-MIR445d	MI0001712	chr5:18245235-18245546(-)	NDNA2TNA_OS (DNA/Tourist)	96.15
osa-MIR809b	MI0005222	chr5:26781125-26781276(-)	STOWAWAY1_OS (DNA/Stowaway)	86.18
osa-MIR819g	MI0005258	chr5:28003948-28004094(+)	STOWAWAY1_OS (DNA/Stowaway)	100.00
osa-MIR815a	MI0005242	chr5:29663360-29663442(+)	WANDERER_OS (DNA/Tourist)	26.51
osa-MIR439h	MI0001698	chr6:1552120-1552218(-)	MuDR4_OS (DNA/MuDR)	88.89
osa-MIR819h	MI0005259	chr6:10052973-10053127(-)	STOWAWAY50_OS (DNA/Stowaway), SZ-66LTR (LTR/Gypsy)	100.00
osa-MIR446	MI0001718	chr6:11975952-11976126(+)	STOWAWAY1_OS (DNA/Stowaway), MIDWAY (DNA)	90.86
osa-MIR811a	MI0005230	chr6:13901553-13901742(+)	TAMI2 (DNA)	100.00
osa-MIR438	MI0001689	chr6:20744968-20745153(+)	OSTE1 (DNA)	1.61
osa-MIR812c	MI0005235	chr6:26259310-26259473(+)	STOWAWAY9_OS (DNA/Stowaway)	100.00
osa-MIR809c	MI0005223	chr6:29694703-29694857(+)	STOWAWAY1_OS (DNA/Stowaway), STOWAWAY47_OS (DNA/Stowaway)	98.06
osa-MIR815c	MI0005244	chr7:5425185-5425287(-)	DITTO-2 (DNA/Tourist)	58.25
osa-MIR439c	MI0001693	chr7:8232129-8232221(+)	MuDR4_OS (DNA/MuDR)	96.77
osa-MIR820b	MI0005264	chr7:13118845-13119035(-)	ENSPM2_OS (DNA/En-Spm)	16.75
osa-MIR809d	MI0005224	chr7:14711300-14711466(+)	STOWAWAY1_OS (DNA/Stowaway), MIDWAY (DNA)	95.81

Table E.1 continued

Name[a]	Accn[b]	Coords[c]	TE[d]	Overlap[e]
osa-MIR821b	MI0005267	chr7:16415531-16415817(+)	OSTE22 (DNA), TNR3_OS (DNA/En-Spm)	100.00
osa-MIR441b	MI0001705	chr7:17310115-17310266(-)	STOWAWAY1_OS (DNA/Stowaway)	87.50
osa-MIR812d	MI0005236	chr7:22393529-22393681(+)	STOWAWAY44_OS (DNA/Stowaway)	100.00
osa-MIR819i	MI0005260	chr7:27791376-27791573(+)	STOWAWAY1_OS (DNA/Stowaway)	75.76
osa-MIR445a	MI0001709	chr7:28117531-28117798(+)	NDNA2TNA_OS (DNA/Tourist)	100.00
osa-MIR818e	MI0005251	chr7:28152738-28152962(-)	STOWAWAY21_OS (DNA/Stowaway)	100.00
osa-MIR806e	MI0005214	chr7:29167591-29167845(+)	TREP215 (DNA/Stowaway)	93.33
osa-MIR531	MI0003204	chr8:1214013-1214093(-)	SC-1_int-int (LTR/Copia)	100.00
osa-MIR439g	MI0001697	chr8:3657884-3657971(+)	MuDR4_OS (DNA/MuDR)	96.59
osa-MIR439e	MI0001695	chr8:14924726-14924823(-)	MuDR4_OS (DNA/MuDR)	91.84
osa-MIR809e	MI0005225	chr8:15317880-15318040(-)	STOWAWAY1_OS (DNA/Stowaway)	97.52
osa-MIR812e	MI0005237	chr8:16268303-16268472(+)	STOWAWAY44_OS (DNA/Stowaway)	100.00
osa-MIR821c	MI0005268	chr8:19792287-19792552(-)	ENSPM3_OS (DNA/En-Spm), OSTE22 (DNA)	100.00
osa-MIR806f	MI0005215	chr8:27832638-27832878(+)	TREP215 (DNA/Stowaway)	99.17
osa-MIR819j	MI0005261	chr8:28081183-28081357(+)	STOWAWAY1_OS (DNA/Stowaway)	85.71
osa-MIR809f	MI0005226	chr9:12000642-12000795(+)	STOWAWAY1_OS (DNA/Stowaway)	97.40
osa-MIR439f	MI0001696	chr9:16962023-16962118(+)	MuDR4_OS (DNA/MuDR)	93.75
osa-MIR811b	MI0005231	chr10:2372014-2372203(+)	TAMI2 (DNA)	100.00
osa-MIR439b	MI0001692	chr10:5338996-5339055(+)	MuDR4_OS (DNA/MuDR)	100.00
osa-MIR820c	MI0005265	chr10:6693845-6694025(+)	ENSPM2_OS (DNA/En-Spm)	14.92
osa-MIR439j	MI0001700	chr10:15479858-15479982(-)	MuDR4_OS (DNA/MuDR)	26.40
osa-MIR819k	MI0005262	chr10:18374459-18374614(-)	STOWAWAY50_OS (DNA/Stowaway)	98.72
osa-MIR816	MI0005245	chr10:21478646-21478722(+)	STOWAWAY47_OS (DNA/Stowaway)	100.00

Table E.1 continued

Name[a]	Accn[b]	Coords[c]	TE[d]	Overlap[e]
osa-MIR806g	MI0005216	chr10:22588399-22588638(+)	TREP215 (DNA/Stowaway)	100.00
osa-MIR811c	MI0005232	chr11:5200383-5200541(-)	TAMI2 (DNA)	100.00
osa-MIR813	MI0005238	chr11:23113437-23113639(+)	NDNA1TNA_OS (DNA/Tourist)	100.00
osa-MIR809g	MI0005227	chr11:25437945-25438096(-)	STOWAWAY1_OS (DNA/Stowaway)	98.68
osa-MIR531	MI0003204	chr11:26423868-26423948(+)	SC-1_int-int (LTR/Copia)	100.00
osa-MIR806h	MI0005217	chr11:28361967-28362237(-)	TREP215 (DNA/Stowaway)	88.19
osa-MIR809h	MI0005228	chr12:5776955-5777088(+)	STOWAWAY1_OS (DNA/Stowaway)	100.00
osa-MIR419	MI0001439	chr12:8234925-8235027(-)	CACTA-I (DNA/En-Spm)	5.83

[a]miRNA name (from miRBase)
[b]miRBase accession number
[c]Genomic coordinates of the miRNA
[d]Name of co-located TE
[e]Percent of miRNA overlapping with TE sequence

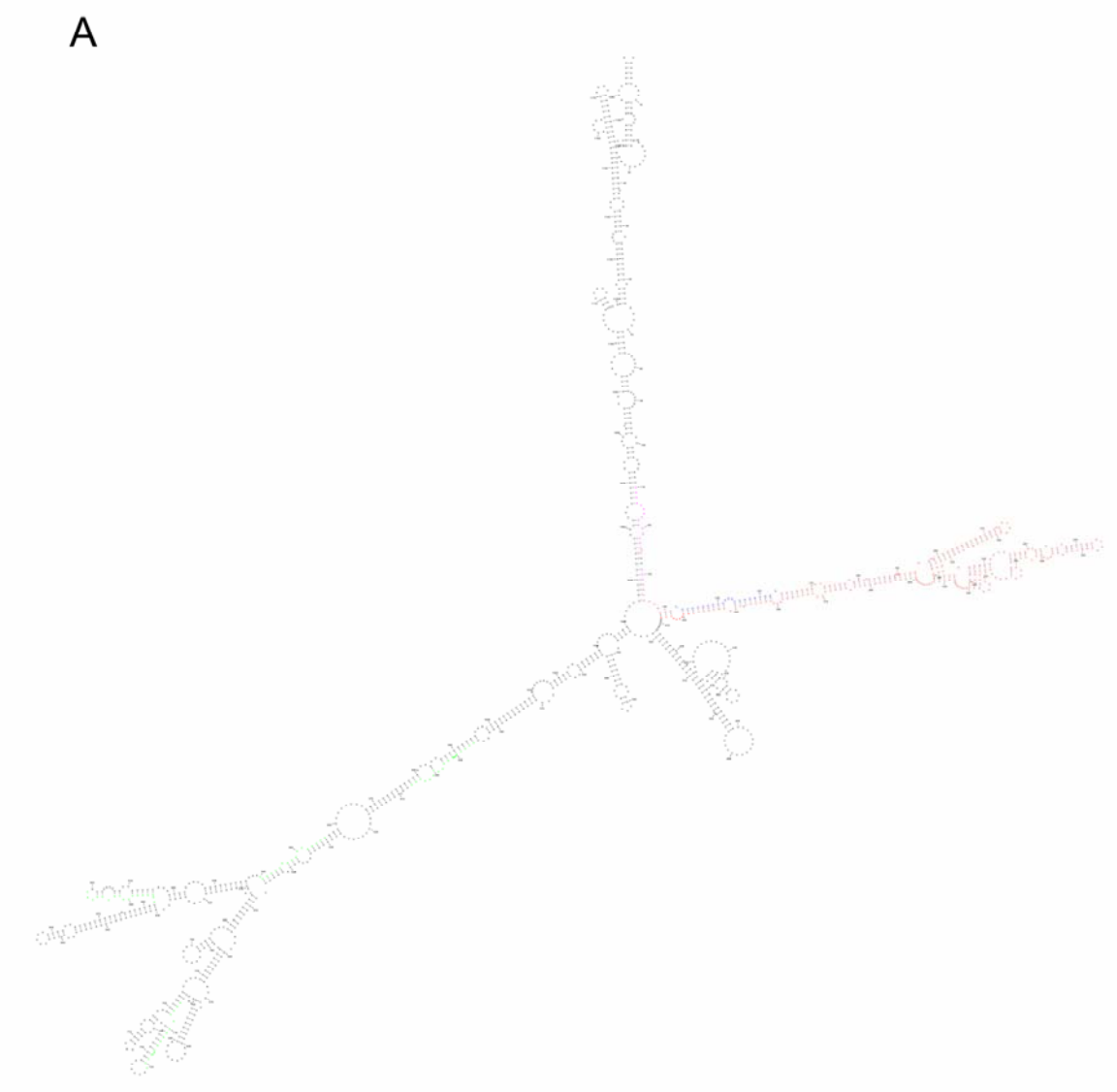

Figure E.1: RNA secondary structures of siRNA-miRNA encoding TE sequences. The predicted secondary structures of TE sequence transcripts encoding both siRNA and miRNA are shown for (A) ath-MIR855 (B) ath-MIR401 (C) ath-MIR416 (D) ath-MIR405b (E) ath-MIR854a (F) osa-MIR439a (G) osa-MIR814a (H) osa-MIR812b (I) osa-MIR814b (J) osa-MIR821a (K) osa-MIR443 (L) osa-MIR420 (M) osa-MIR442 (N) osa-MIR531 (O) osa-MIR821c (P) osa-MIR439b (Q) osa-MIR531 (R) osa-MIR821b. The miRNA stem-loop region, miRNA mature sequence, miRNA signature sequence and siRNA signature sequence are shown in red, blue, pink and green, respectively.

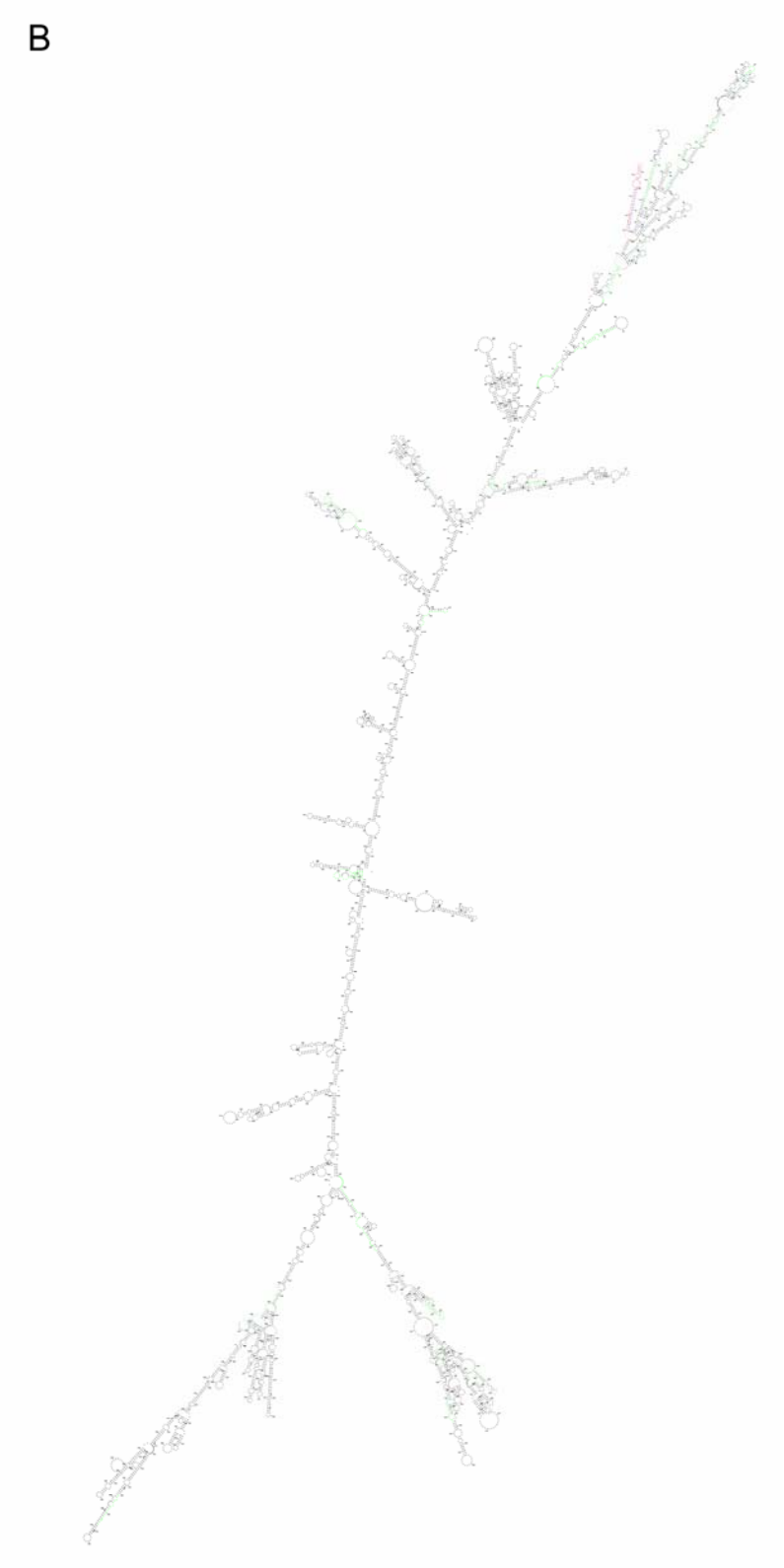

Figure E.1 continued

C

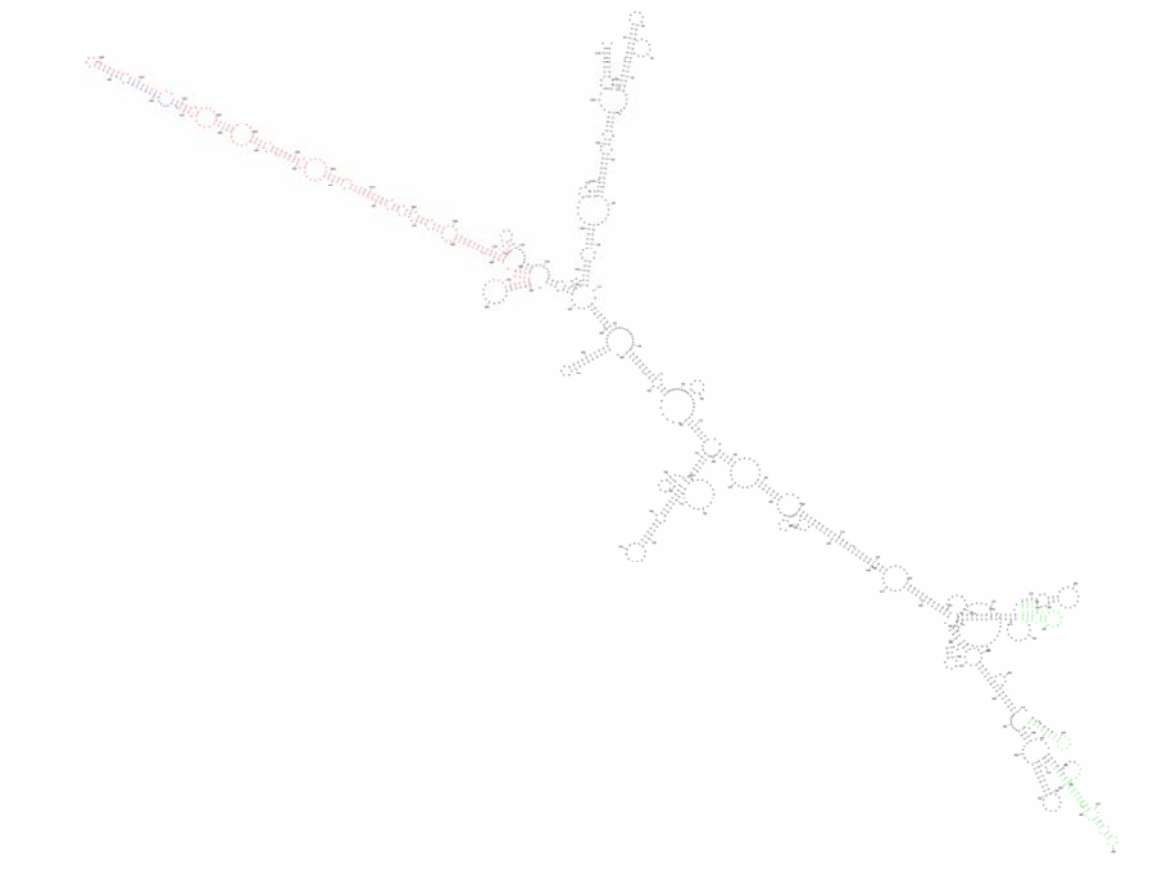

Figure E.1 continued

D

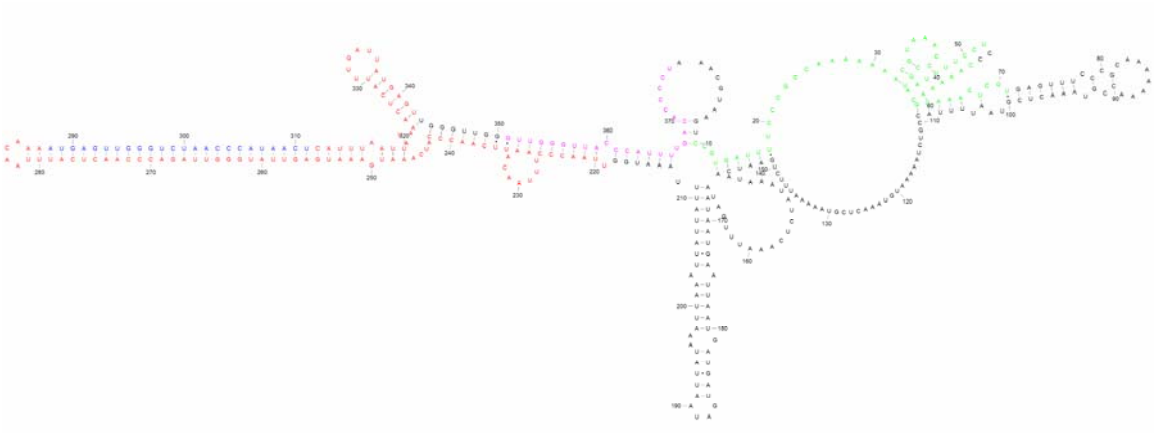

Figure E.1 continued

E

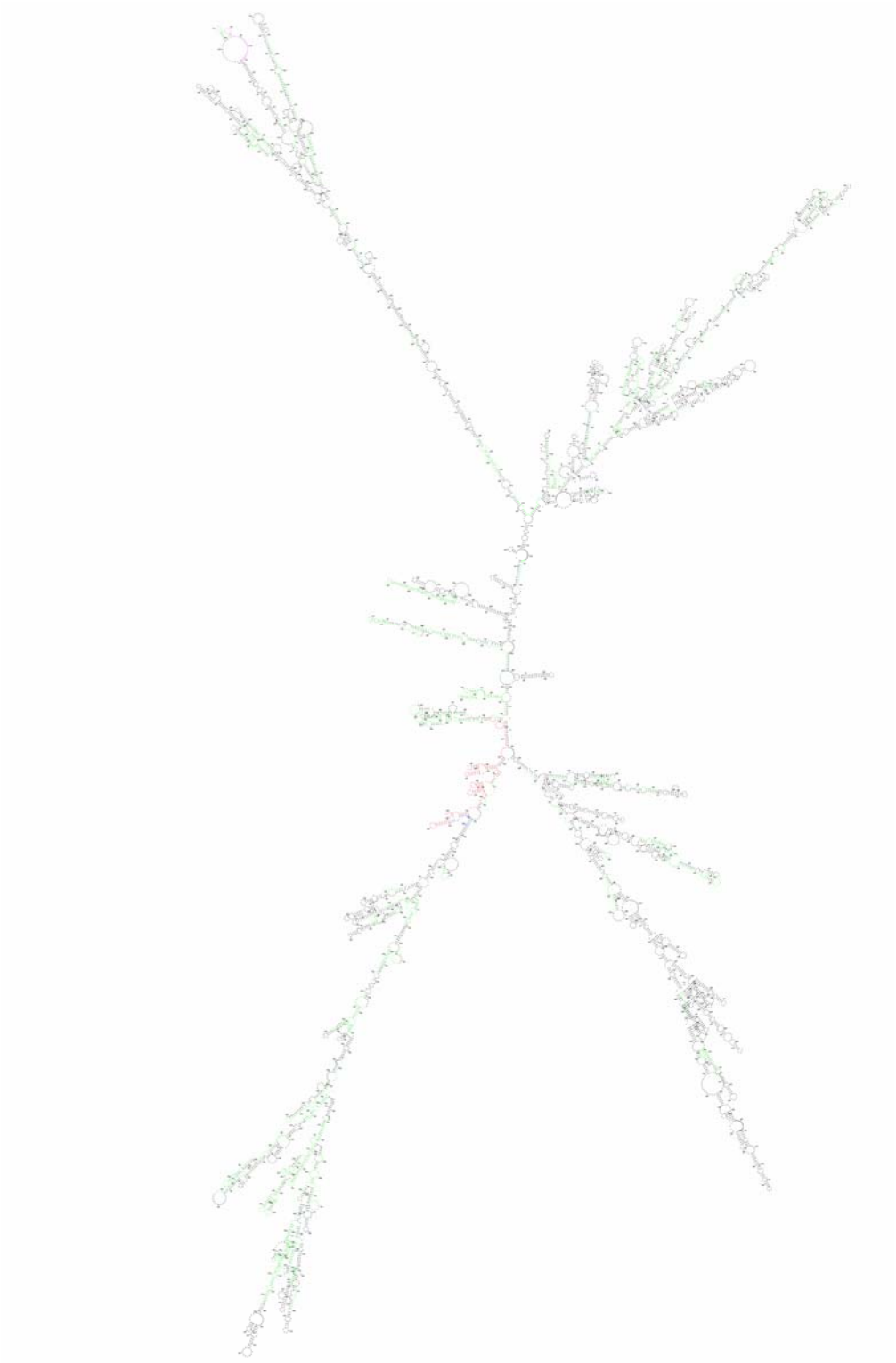

Figure E.1 continued

F

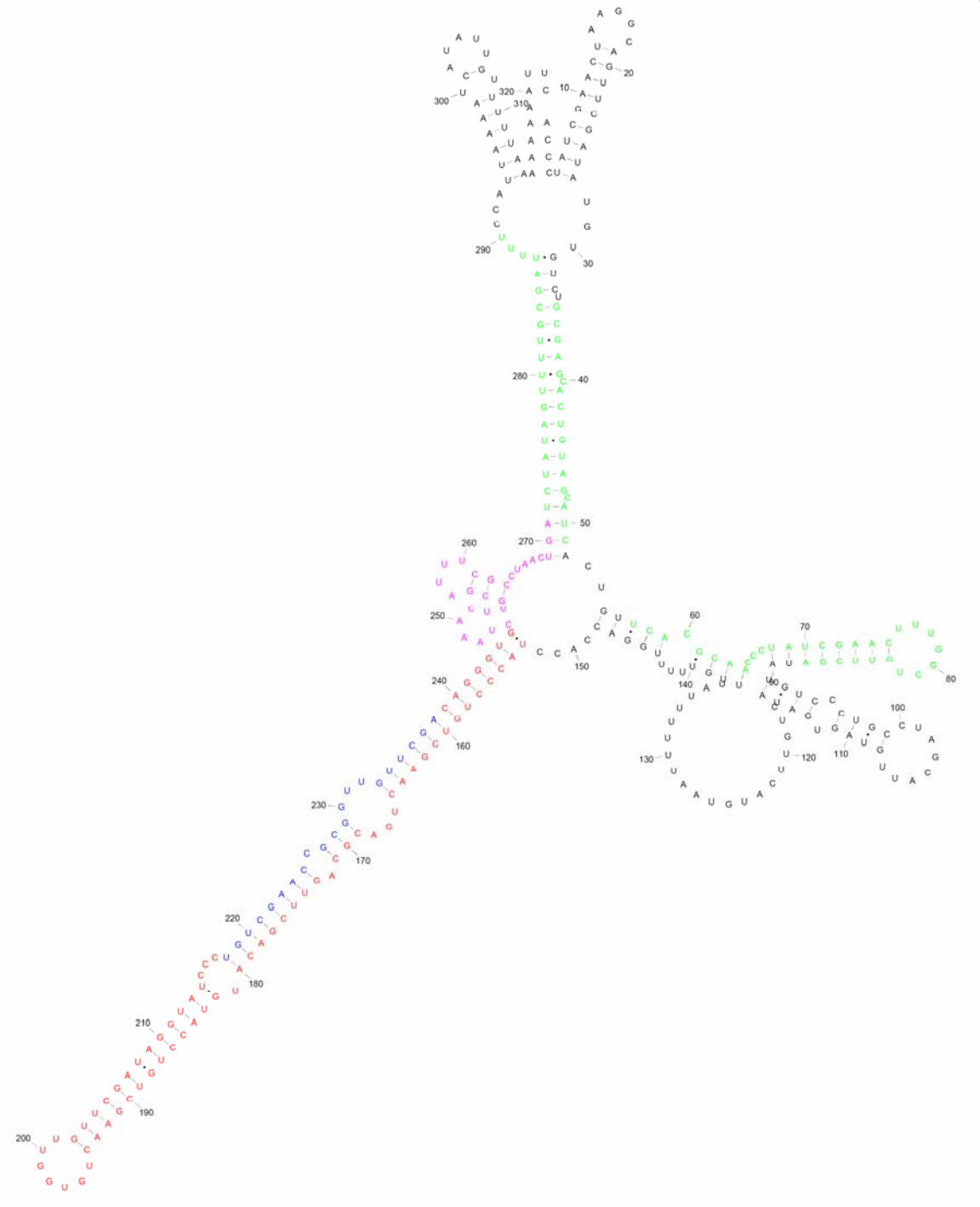

Figure E.1 continued

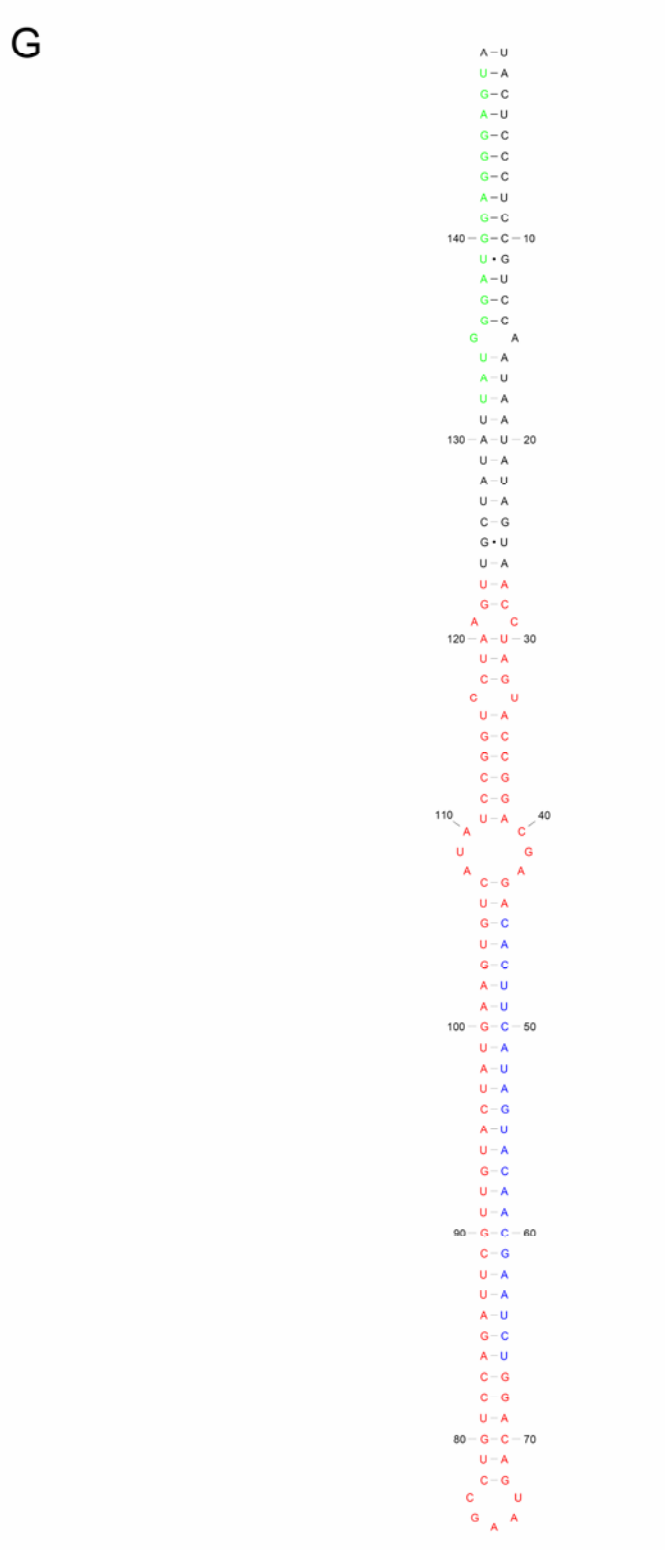

Figure E.1 continued

H

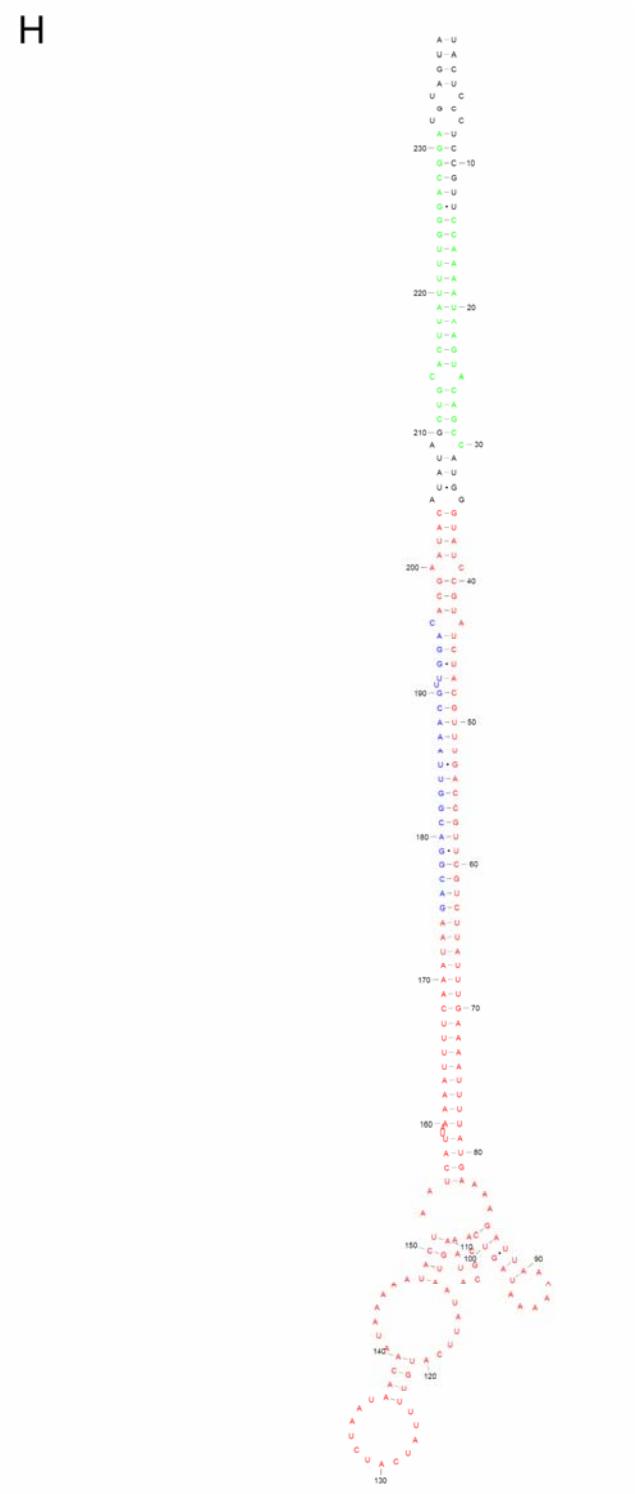

Figure E.1 continued

Figure E.1 continued

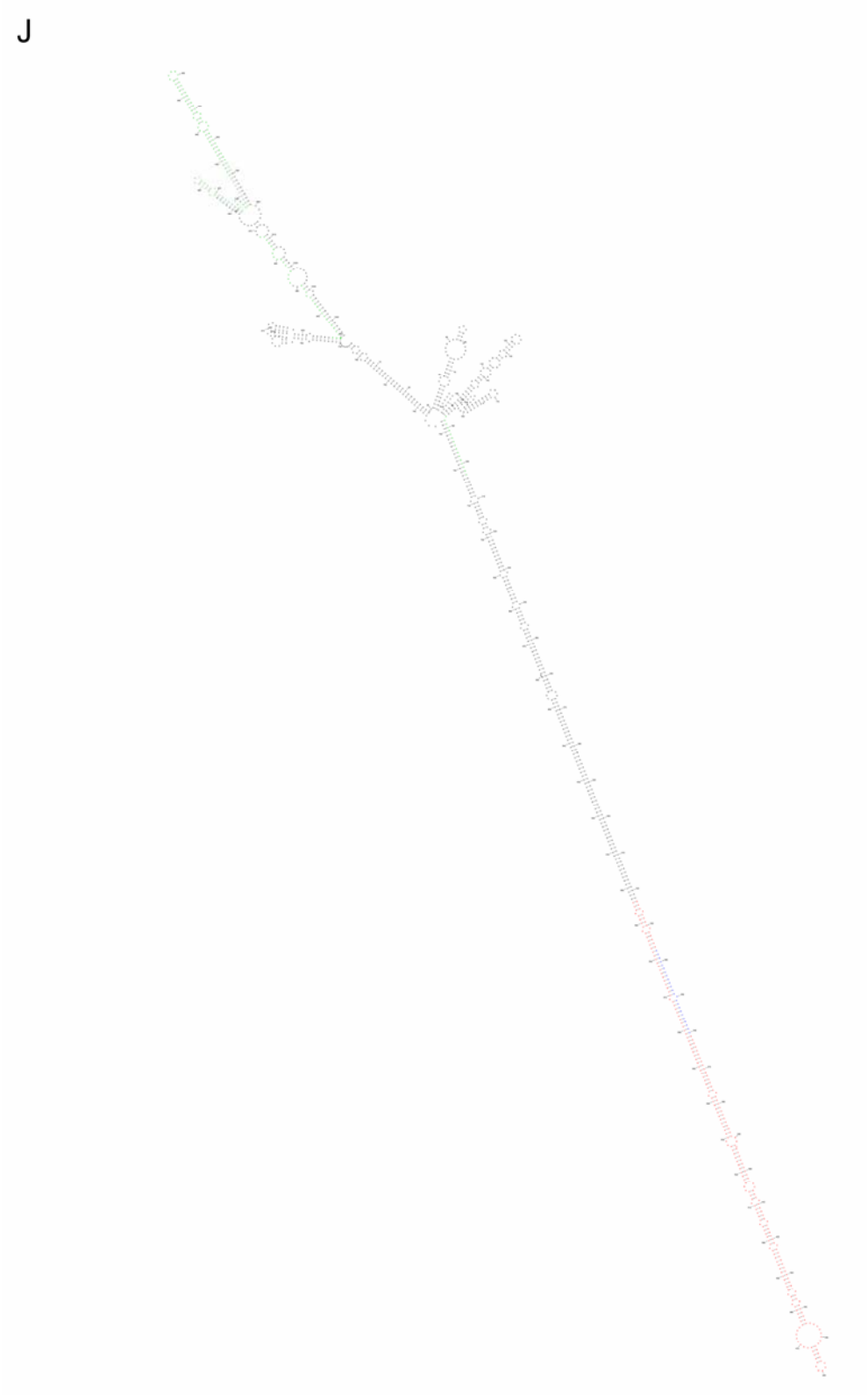

Figure E.1 continued

Figure E.1 continued

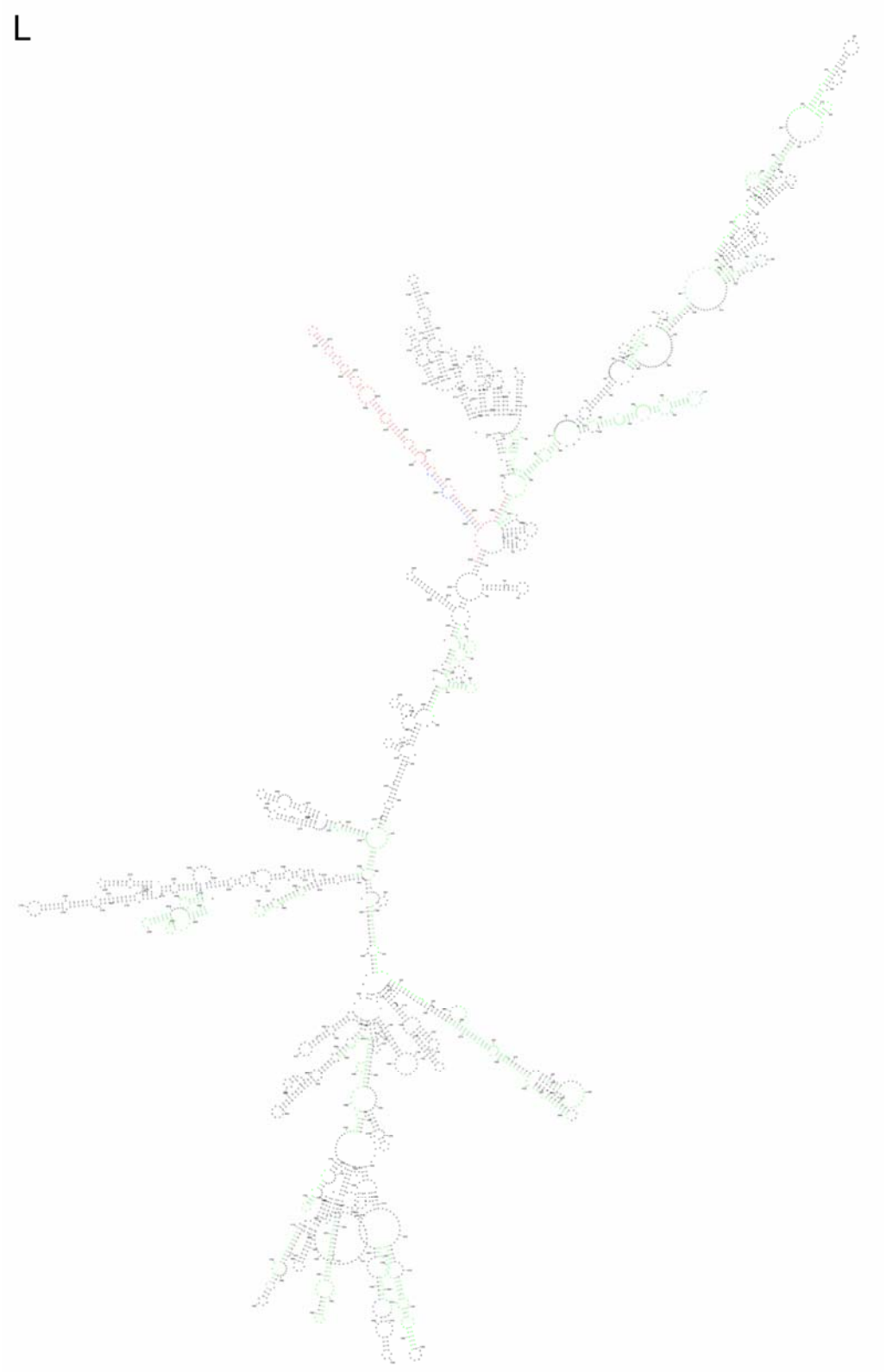

Figure E.1 continued

Figure E.1 continued

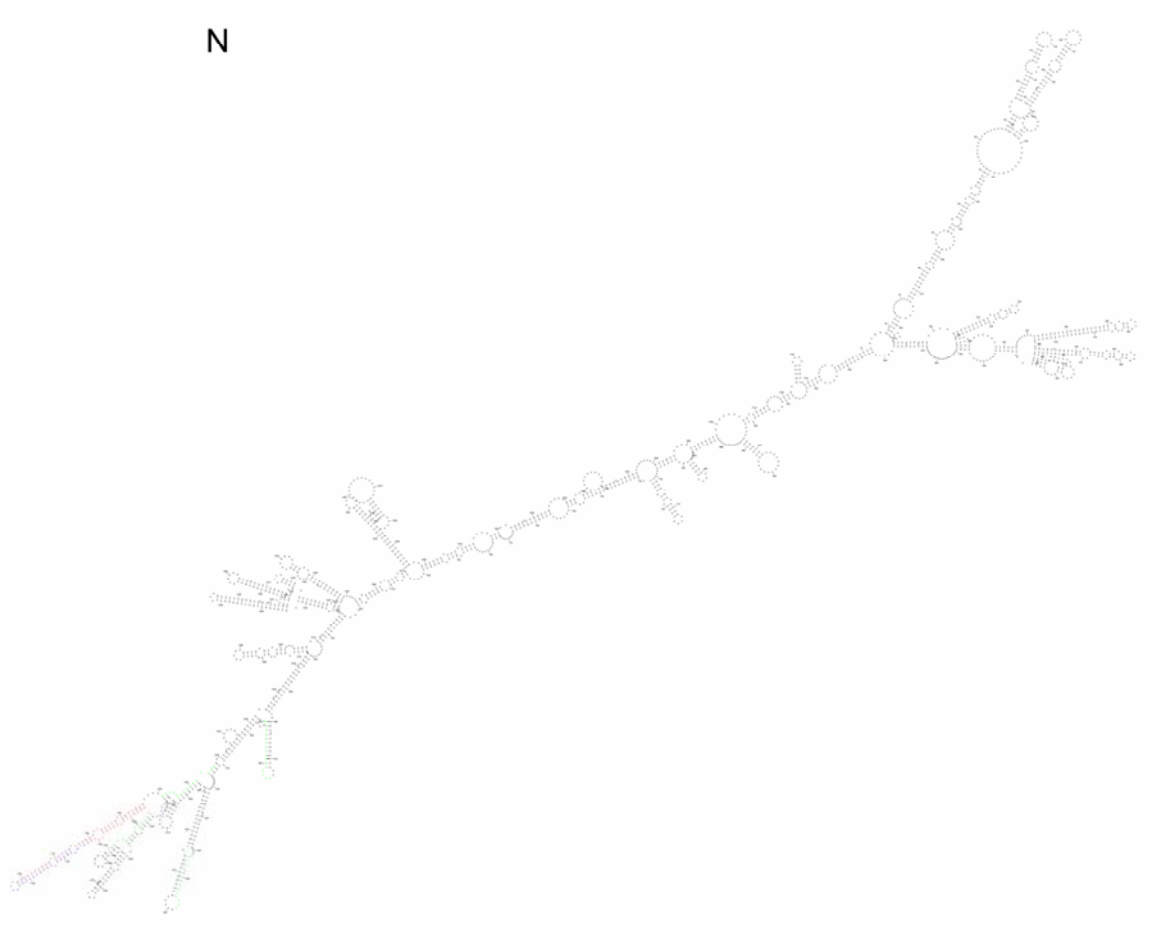

Figure E.1 continued

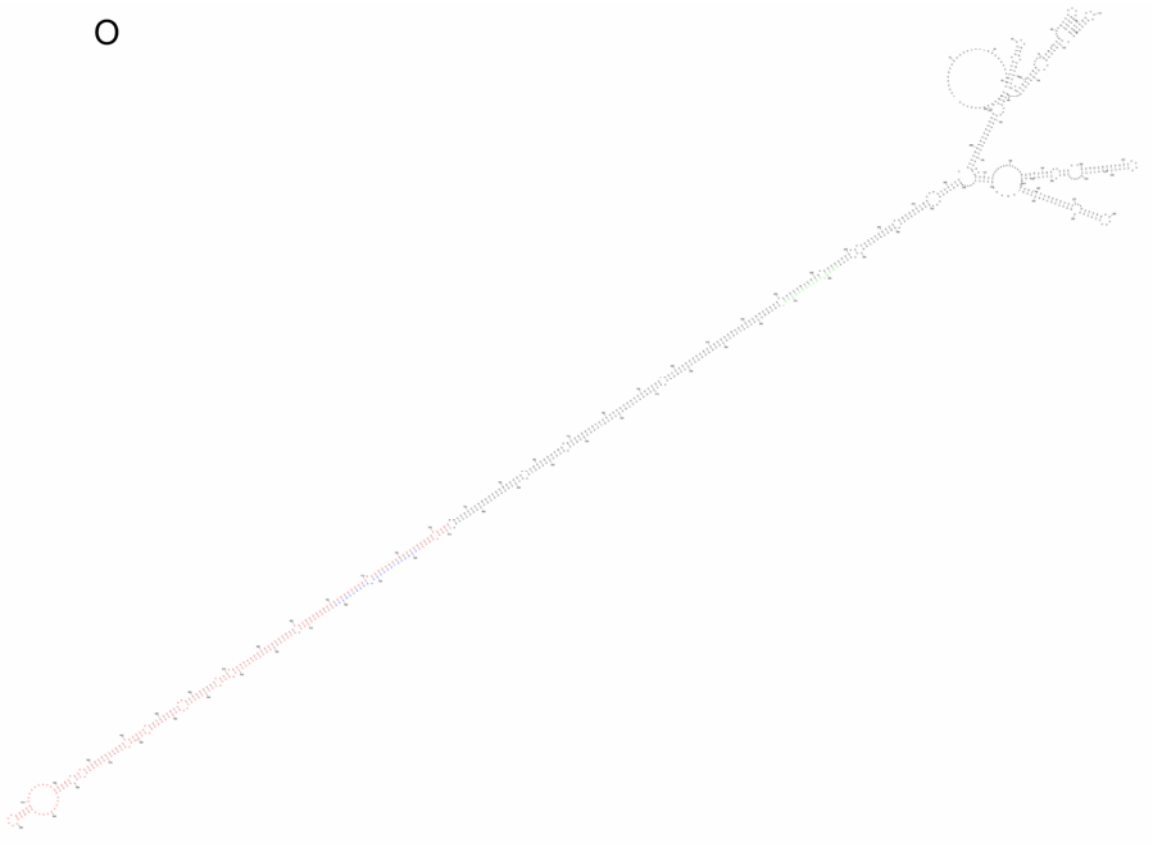

Figure E.1 continued

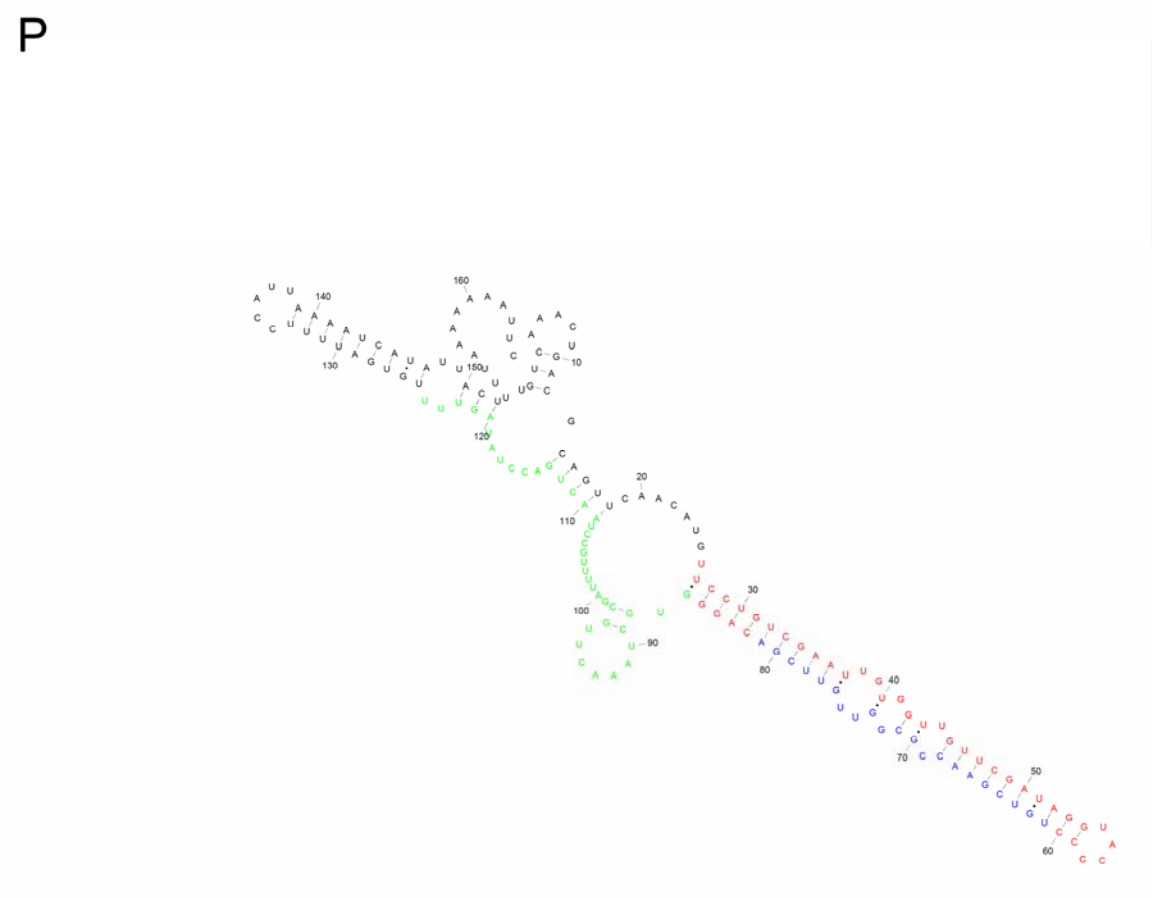

Figure E.1 continued

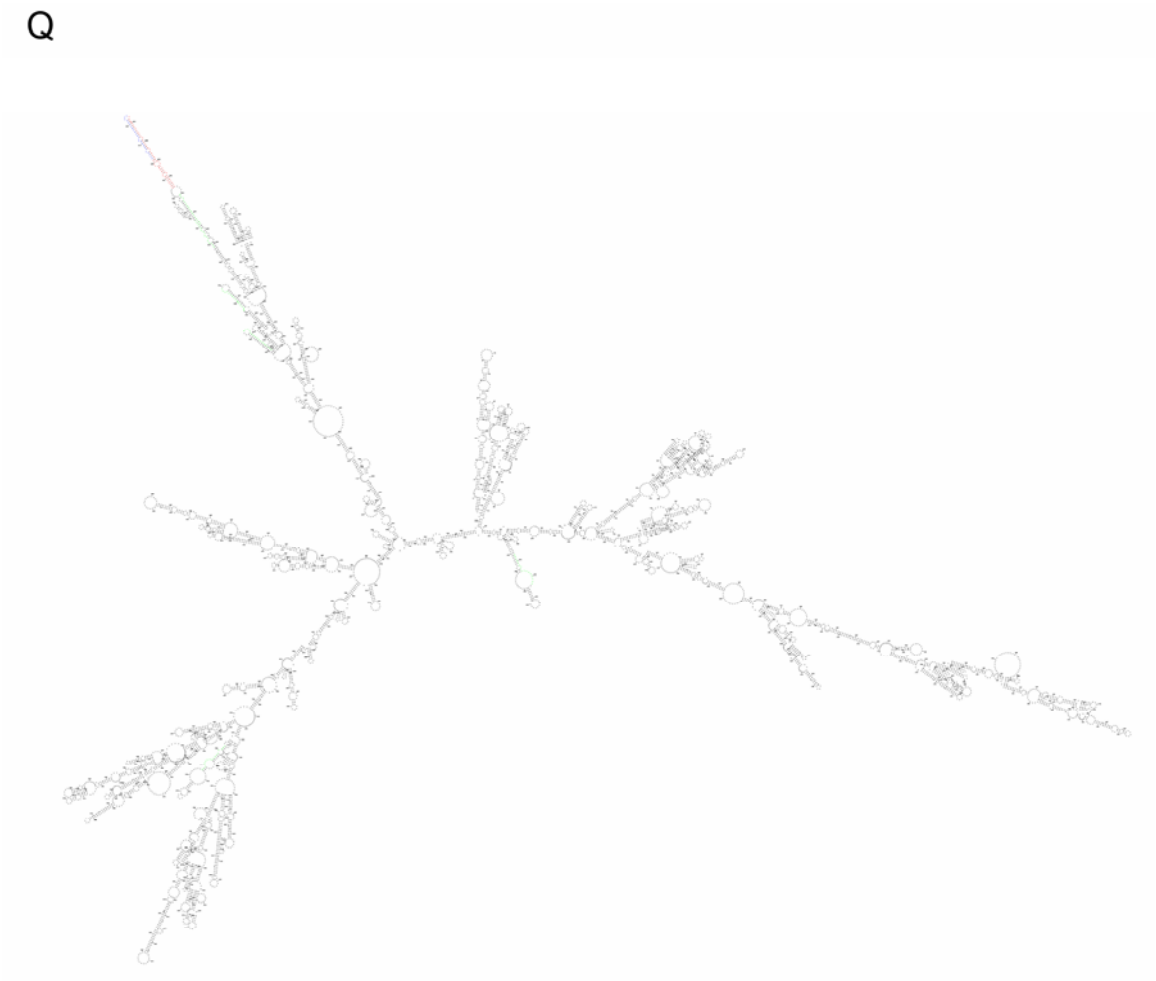

Figure E.1 continued

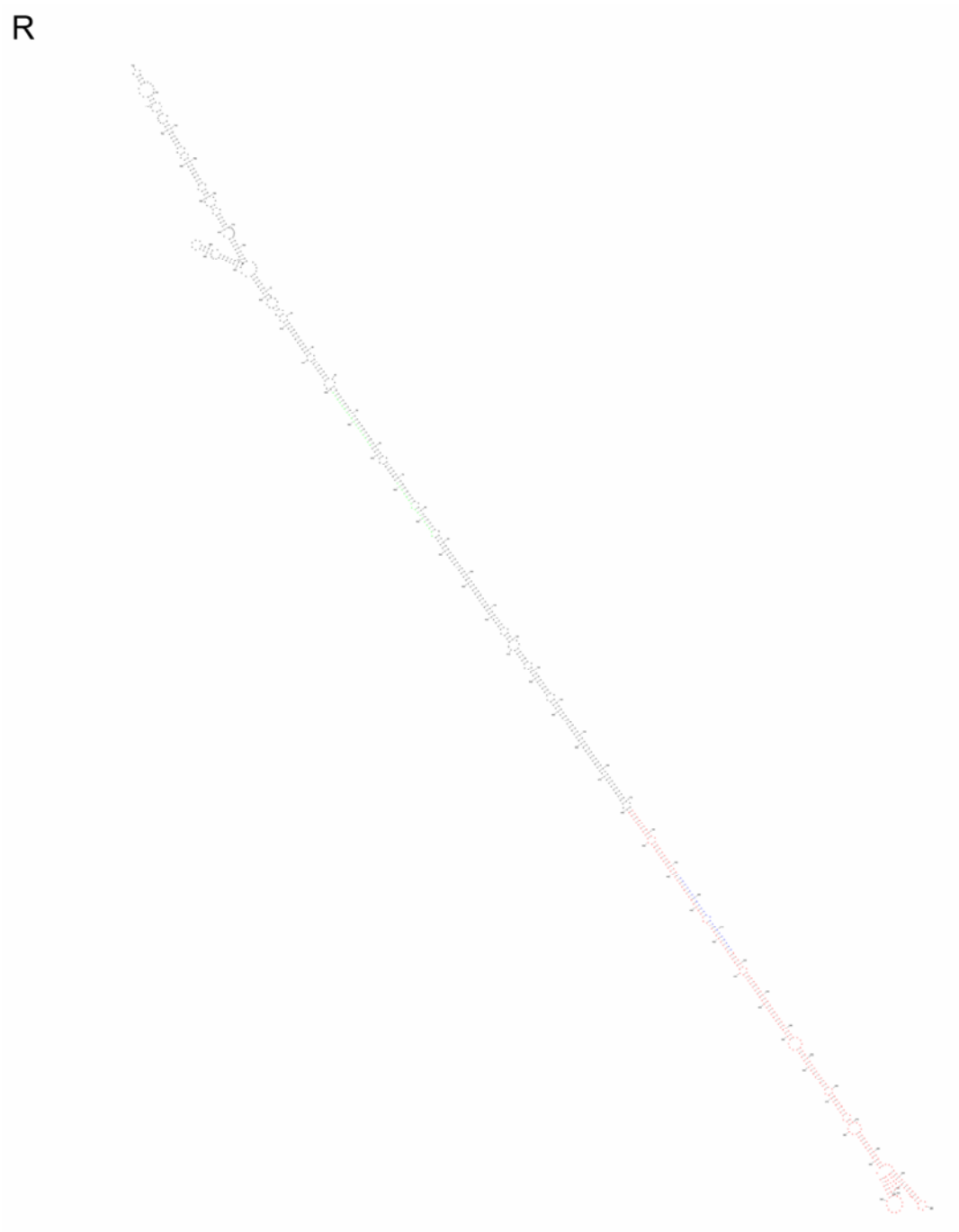

Figure E.1 continued

PUBLICATIONS

1. Piriyapongsa, J., and I. K. Jordan. 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS ONE 2:e203.

2. Piriyapongsa, J., L. Mariño-Ramírez, and I. K. Jordan. 2007a. Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323-1337.

3. Piriyapongsa, J., N. Polavarapu, M. Borodovsky, and J. McDonald. 2007b. Exonization of the LTR transposable elements in human genome. BMC genomics 8:291.

4. Piriyapongsa, J., M. T. Rutledge, S. Patel, M. Borodovsky, and I. K. Jordan. 2007c. Evaluating the protein coding potential of exonized transposable element sequences. Biology direct 2:31.

5. Piriyapongsa, J., and I. K. Jordan. 2008. Dual coding of siRNAs and miRNAs by plant transposable elements. RNA 14:814-821.

REFERENCES

ABOUELHODA, M. I., S. KURTZ and E. OHLEBUSCH, 2004 Replacing suffix trees with enhanced suffix arrays. J Discrete Algorithm **2:** 53-86.

ACKERMAN, H., I. UDALOVA, J. HULL and D. KWIATKOWSKI, 2002 Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. Mol Biol Evol **19:** 884-890.

ADAMS, M. D., S. E. CELNIKER, R. A. HOLT, C. A. EVANS, J. D. GOCAYNE *et al.*, 2000 The genome sequence of Drosophila melanogaster. Science **287:** 2185-2195.

AGRAWAL, A., Q. M. EASTMAN and D. G. SCHATZ, 1998 Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature **394:** 744-751.

ALLEN, E., Z. XIE, A. M. GUSTAFSON, G. H. SUNG, J. W. SPATAFORA *et al.*, 2004 Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat Genet **36:** 1282-1290.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J Mol Biol **215:** 403-410.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-3402.

ALVAREZ-GARCIA, I., and E. A. MISKA, 2005 MicroRNA functions in animal development and human disease. Development **132:** 4653-4662.

AMBROS, V., 2004 The functions of animal microRNAs. Nature **431:** 350-355.

AMBROS, V., B. BARTEL, D. P. BARTEL, C. B. BURGE, J. C. CARRINGTON *et al.*, 2003 A uniform system for microRNA annotation. Rna **9:** 277-279.

APARICIO, S., J. CHAPMAN, E. STUPKA, N. PUTNAM, J. M. CHIA *et al.*, 2002 Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science **297:** 1301-1310.

APWEILER, R., T. K. ATTWOOD, A. BAIROCH, A. BATEMAN, E. BIRNEY *et al.*, 2001 The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res **29:** 37-40.

ARAVIN, A. A., M. LAGOS-QUINTANA, A. YALCIN, M. ZAVOLAN, D. MARKS *et al.*, 2003 The small RNA profile during Drosophila melanogaster development. Dev Cell **5:** 337-350.

ARONOW, B. J., R. N. SILBIGER, M. R. DUSING, J. L. STOCK, K. L. YAGER *et al.*, 1992 Functional analysis of the human adenosine deaminase gene thymic regulatory region and its ability to generate position-independent transgene expression. Mol Cell Biol **12:** 4170-4185.

ARRANZ, V., M. KRESS and M. ERNOULT-LANGE, 1994 The gene encoding the MOK-2 zinc-finger protein: characterization of its promoter and negative regulation by mouse Alu type-2 repetitive elements. Gene **149:** 293-298.

ARTEAGA-VAZQUEZ, M., J. CABALLERO-PEREZ and J. P. VIELLE-CALZADA, 2006 A family of microRNAs present in plants and animals. Plant Cell **18:** 3355-3369.

ASHBURNER, M., C. A. BALL, J. A. BLAKE, D. BOTSTEIN, H. BUTLER *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25:** 25-29.

BABAK, T., W. ZHANG, Q. MORRIS, B. J. BLENCOWE and T. R. HUGHES, 2004 Probing microRNAs with microarrays: tissue specificity and functional inference. Rna **10:** 1813-1819.

BABICH, V., N. AKSENOV, V. ALEXEENKO, S. L. OEI, G. BUCHLOW *et al.*, 1999 Association of some potential hormone response elements in human genes with the Alu family repeats. Gene **239:** 341-349.

BANIAHMAD, A., M. MULLER, C. STEINER and R. RENKAWITZ, 1987 Activity of two different silencer elements of the chicken lysozyme gene can be compensated by enhancer elements. EMBO J **6:** 2297-2303.

BANVILLE, D., and Y. BOIE, 1989 Retroviral long terminal repeat is the promoter of the gene encoding the tumor-associated calcium-binding protein oncomodulin in the rat. J Mol Biol **207:** 481-490.

BARAD, O., E. MEIRI, A. AVNIEL, R. AHARONOV, A. BARZILAI *et al.*, 2004 MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. Genome Res **14:** 2486-2494.

BARTEL, D. P., 2004 MicroRNAs: genomics, biogenesis, mechanism, and function. Cell **116:** 281-297.

BAUMRUKER, T., C. GEHE and I. HORAK, 1988 Insertion of a retrotransposon within the 3' end of a mouse gene provides a new functional polyadenylation signal. Nucleic Acids Res **16:** 7241-7251.

BEJERANO, G., C. B. LOWE, N. AHITUV, B. KING, A. SIEPEL *et al.*, 2006 A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature **441:** 87-90.

BENNETZEN, J. L., 2000 Transposable element contributions to plant gene and genome evolution. Plant Mol Biol **42:** 251-269.

BENTWICH, I., A. AVNIEL, Y. KAROV, R. AHARONOV, S. GILAD *et al.*, 2005 Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet **37:** 766-770.

BEREZIKOV, E., E. CUPPEN and R. H. PLASTERK, 2006 Approaches to microRNA discovery. Nat Genet **38 Suppl:** S2-7.

BEREZIKOV, E., V. GURYEV, J. VAN DE BELT, E. WIENHOLDS, R. H. PLASTERK *et al.*, 2005 Phylogenetic shadowing and computational identification of human microRNA genes. Cell **120:** 21-24.

BERMAN, H. M., J. WESTBROOK, Z. FENG, G. GILLILAND, T. N. BHAT *et al.*, 2000 The Protein Data Bank. Nucleic Acids Res **28:** 235-242.

BERNSTEIN, E., A. A. CAUDY, S. M. HAMMOND and G. J. HANNON, 2001 Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature **409:** 363-366.

BERTUCCI, F., S. SALAS, S. EYSTERIES, V. NASSER, P. FINETTI *et al.*, 2004 Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. Oncogene **23:** 1377-1391.

BEST, S., P. LE TISSIER, G. TOWERS and J. P. STOYE, 1996 Positional cloning of the mouse retrovirus restriction gene Fv1. Nature **382:** 826-829.

BI, S., O. GAVRILOVA, D. W. GONG, M. M. MASON and M. REITMAN, 1997 Identification of a placental enhancer for the human leptin gene. J Biol Chem **272:** 30583-30588.

BIECHE, I., A. LAURENT, I. LAURENDEAU, L. DURET, Y. GIOVANGRANDI *et al.*, 2003 Placenta-specific INSL4 expression is mediated by a human endogenous retrovirus element. Biol Reprod **68:** 1422-1429.

BIEMONT, C., and C. VIEIRA, 2005 What transposable elements tell us about genome organization and evolution: the case of Drosophila. Cytogenet Genome Res **110:** 25-34.

BIEMONT, C., and C. VIEIRA, 2006 Genetics: junk DNA as an evolutionary force. Nature **443:** 521-524.

BLACKBURN, E. H., 1991 Structure and function of telomeres. Nature **350:** 569-573.

BLAISE, S., N. DE PARSEVAL, L. BENIT and T. HEIDMANN, 2003 Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci U S A **100:** 13013-13018.

BLAISE, S., N. DE PARSEVAL and T. HEIDMANN, 2005 Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes. Retrovirology **2:** 19.

BLANCHETTE, M., W. J. KENT, C. RIEMER, L. ELNITSKI, A. F. SMIT *et al.*, 2004 Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res **14:** 708-715.

BLOND, J. L., F. BESEME, L. DURET, O. BOUTON, F. BEDIN *et al.*, 1999 Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. J Virol **73:** 1175-1185.

BOECKMANN, B., A. BAIROCH, R. APWEILER, M. C. BLATTER, A. ESTREICHER *et al.*, 2003 The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res **31:** 365-370.

BOGUSKI, M. S., T. M. LOWE and C. M. TOLSTOSHEV, 1993 dbEST--database for "expressed sequence tags". Nat Genet **4:** 332-333.

BORCHERT, G. M., W. LANIER and B. L. DAVIDSON, 2006 RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol **13:** 1097-1101.

BORODOVSKY, M., and J. MCININCH, 1993 GeneMark: parallel gene recognition for both DNA strands. Computers and Chemistry **17:** 123-133.

BOWEN, N. J., and I. K. JORDAN, 2007 Exaptation of protein coding sequences from transposable elements. Genome Dynamics **3:** 131-146.

BOWEN, N. J., I. K. JORDAN, J. A. EPSTEIN, V. WOOD and H. L. LEVIN, 2003 Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements from the complete genome sequence of Schizosaccharomyces pombe. Genome Res **13:** 1984-1997.

BRANDT, J., S. SCHRAUTH, A. M. VEITH, A. FROSCHAUER, T. HANEKE *et al.*, 2005a Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene **345:** 101-111.

BRANDT, J., A. M. VEITH and J. N. VOLFF, 2005b A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes. Cytogenet Genome Res **110:** 307-317.

BRENNECKE, J., A. A. ARAVIN, A. STARK, M. DUS, M. KELLIS *et al.*, 2007 Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell **128:** 1089-1103.

BRENNECKE, J., D. R. HIPFNER, A. STARK, R. B. RUSSELL and S. M. COHEN, 2003 bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell **113:** 25-36.

BRENNER, S., M. JOHNSON, J. BRIDGHAM, G. GOLDA, D. H. LLOYD *et al.*, 2000a Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol **18:** 630-634.

BRENNER, S., S. R. WILLIAMS, E. H. VERMAAS, T. STORCK, K. MOON *et al.*, 2000b In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. Proc Natl Acad Sci U S A **97:** 1665-1670.

BRINI, A. T., G. M. LEE and J. P. KINET, 1993 Involvement of Alu sequences in the cell-specific regulation of transcription of the gamma chain of Fc and T cell receptors. J Biol Chem **268:** 1355-1361.

BRITTEN, R., 2006 Transposable elements have contributed to thousands of human proteins. Proc Natl Acad Sci U S A **103:** 1798-1803.

BRITTEN, R. J., 1996 DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci U S A **93:** 9374-9377.

BRITTEN, R. J., 1997 Mobile elements inserted in the distant past have taken on important functions. Gene **205:** 177-182.

BRITTEN, R. J., and E. H. DAVIDSON, 1969 Gene regulation for higher cells: a theory. Science **165:** 349-357.

BROSIUS, J., 1991 Retroposons--seeds of evolution. Science **251:** 753.

BROSIUS, J., 1999 Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica **107:** 209-238.

BROSIUS, J., and S. J. GOULD, 1992 On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci U S A **89:** 10706-10710.

BUCHON, N., and C. VAURY, 2006 RNAi: a defensive RNA-silencing against viruses and transposable elements. Heredity **96:** 195-202.

BUNDOCK, P., and P. HOOYKAAS, 2005 An Arabidopsis hAT-like transposase is essential for plant development. Nature **436:** 282-284.

BUREAU, T. E., and S. R. WESSLER, 1992 Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell **4:** 1283-1294.

BUREAU, T. E., and S. R. WESSLER, 1994 Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell **6:** 907-916.

CAMPILLOS, M., T. DOERKS, P. K. SHAH and P. BORK, 2006 Computational characterization of multiple Gag-like human proteins. Trends Genet **22:** 585-589.

CARROLL, D., D. KNUTZON and J. GARRETT, 1989 Transposable elements in Xenopus species, pp. 567-574 in *Mobile DNA*, edited by M. HOWE and D. BERG. American Society Microbiolgy, Washington DC.

CERUTTI, H., and J. A. CASAS-MOLLANO, 2006 On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet **50:** 81-99.

CHANG-YEH, A., D. E. MOLD and R. C. HUANG, 1991 Identification of a novel murine IAP-promoted placenta-expressed gene. Nucleic Acids Res **19:** 3667-3672.

CHARLESWORTH, B., P. SNIEGOWSKI and W. STEPHAN, 1994 The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371:** 215-220.

CHEN, C. Z., L. LI, H. F. LODISH and D. P. BARTEL, 2004 MicroRNAs modulate hematopoietic lineage differentiation. Science **303:** 83-86.

CHEN, H. J., K. CARR, R. E. JEROME and H. J. EDENBERG, 2002 A retroviral repetitive element confers tissue-specificity to the human alcohol dehydrogenase 1C (ADH1C) gene. DNA Cell Biol **21:** 793-801.

CHENG, J., P. KAPRANOV, J. DRENKOW, S. DIKE, S. BRUBAKER *et al.*, 2005 Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science **308:** 1149-1154.

CLAVERIE, J. M., 2005 Fewer genes, more noncoding RNA. Science **309:** 1529-1530.

CONLEY, A. B., W. J. MILLER and I. K. JORDAN, 2008 Human cis natural antisense transcripts initiated by transposable elements. Trends Genet **24:** 53-56.

CONTE, C., B. DASTUGUE and C. VAURY, 2002 Promoter competition as a mechanism of transcriptional interference mediated by retrotransposons. Embo J **21:** 3908-3916.

CORDAUX, R., S. UDIT, M. A. BATZER and C. FESCHOTTE, 2006 Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci U S A **103:** 8101-8106.

COVEY, S. N., N. S. AL-KAFF, A. LÁNGARA and D. S. TURNER, 1997 Plants combat infection by gene silencing. Nature **385:** 781-782.

COWAN, R. K., D. R. HOEN, D. J. SCHOEN and T. E. BUREAU, 2005 MUSTANG is a novel family of domesticated transposase genes found in diverse angiosperms. Mol Biol Evol **22:** 2084-2089.

CROLLIUS, H., O. JAILLON, C. DASILVA, C. OZOUF-COSTAZ, C. FIZAMES *et al.*, 2000 Characterization and repeat analysis of the compact genome of the freshwater pufferfish Tetraodon nigroviridis. Genome Res **10:** 939-949.

CRUVEILLER, S., O. CLAY, K. JABBARI and G. BERNARDI, 2007 Simple proteomic checks for detecting noncoding RNA. Proteomics **7:** 361-363.

CULLEN, B. R., 2002 RNA interference: antiviral defense and genetic tool. Nat Immunol **3:** 597-599.

CUMMINS, J. M., Y. HE, R. J. LEARY, R. PAGLIARINI, L. A. DIAZ, JR. *et al.*, 2006 The colorectal microRNAome. Proc Natl Acad Sci U S A **103:** 3687-3692.

DAGAN, T., R. SOREK, E. SHARON, G. AST and D. GRAUR, 2004 AluGene: a database of Alu elements incorporated within protein-coding genes. Nucleic Acids Res **32:** D489-492.

DALMAU, J., S. H. GULTEKIN, R. VOLTZ, R. HOARD, T. DESCHAMPS *et al.*, 1999 Ma1, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders. Brain **122 (Pt 1):** 27-39.

DASILVA, C., H. HADJI, C. OZOUF-COSTAZ, S. NICAUD, O. JAILLON *et al.*, 2002 Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the Tetraodon nigroviridis genome. Proc Natl Acad Sci U S A **99:** 13636-13641.

DASKALOVA, E., V. BAEV, V. RUSINOV and I. MINKOV, 2006 3'UTR-located Alu elements: donors of potetial miRNA target sites and mediators of network miRNA-based regulatory interactions. Evol Bioinform **2:** 99-116.

DE CARVALHO, F., G. GHEYSEN, S. KUSHNIR, M. VAN MONTAGU, D. INZE *et al.*, 1992 Suppression of beta-1,3-glucanase transgene expression in homozygous plants. Embo J **11:** 2595-2602.

DEWANNIEUX, M., C. ESNAULT and T. HEIDMANN, 2003 LINE-mediated retrotransposition of marked Alu sequences. Nat Genet **35:** 41-48.

DI CRISTOFANO, A., M. STRAZZULLO, L. LONGO and G. LA MANTIA, 1995 Characterization and genomic mapping of the ZNF80 locus: expression of this zinc-finger gene is driven by a solitary LTR of ERV9 endogenous retroviral family. Nucleic Acids Res **23:** 2823-2830.

DLAKIC, M., 2002 A model of the replication fork blocking protein Fob1p based on the catalytic core domain of retroviral integrases. Protein Sci **11:** 1274-1277.

DOENCH, J. G., C. P. PETERSEN and P. A. SHARP, 2003 siRNAs can function as miRNAs. Genes Dev **17:** 438-442.

DOMANSKY, A. N., E. P. KOPANTZEV, E. V. SNEZHKOV, Y. B. LEBEDEV, C. LEIB-MOSCH *et al.*, 2000 Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. FEBS Lett **472:** 191-195.

DONNELLY, S. R., T. E. HAWKINS and S. E. MOSS, 1999 A conserved nuclear element with a role in mammalian gene regulation. Hum Mol Genet **8:** 1723-1728.

DOOLITTLE, W. F., and C. SAPIENZA, 1980 Selfish genes, the phenotype paradigm and genome evolution. Nature **284:** 601-603.

DUMOUTIER, L., D. LEJEUNE, D. COLAU and J. C. RENAULD, 2001 Cloning and characterization of IL-22 binding protein, a natural antagonist of IL-10-related T cell-derived inducible factor/IL-22. J Immunol **166:** 7090-7095.

DUNN, C. A., P. MEDSTRAND and D. L. MAGER, 2003 An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. Proc Natl Acad Sci U S A **100:** 12841-12846.

DUPRESSOIR, A., G. MARCEAU, C. VERNOCHET, L. BENIT, C. KANELLOPOULOS *et al.*, 2005 Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci U S A **102:** 725-730.

EDDY, S. R., 1996 Hidden Markov models. Curr Opin Struct Biol **6:** 361-365.

EDDY, S. R., 1998 Profile hidden Markov models. Bioinformatics **14:** 755-763.

EICKBUSH, T. H., 1997 Telomerase and retrotransposons: which came first? Science **277:** 911-912.

EMI, M., A. HORII, N. TOMITA, T. NISHIDE, M. OGAWA *et al.*, 1988 Overlapping two genes in human DNA: a salivary amylase gene overlaps with a gamma-actin pseudogene that carries an integrated human endogenous retroviral DNA. Gene **62:** 229-235.

ENRIGHT, A. J., B. JOHN, U. GAUL, T. TUSCHL, C. SANDER *et al.*, 2003 MicroRNA targets in Drosophila. Genome Biol **5:** R1.

ESPOSITO, T., F. GIANFRANCESCO, A. CICCODICOLA, L. MONTANINI, S. MUMM *et al.*, 1999 A novel pseudoautosomal human gene encodes a putative protein similar to Ac-like transposases. Hum Mol Genet **8:** 61-67.

FARH, K. K., A. GRIMSON, C. JAN, B. P. LEWIS, W. K. JOHNSTON *et al.*, 2005 The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science **310:** 1817-1821.

FESCHOTTE, C., and C. MOUCHES, 2000 Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. Mol Biol Evol **17:** 730-737.

FESCHOTTE, C., and E. J. PRITHAM, 2005 Non-mammalian c-integrases are encoded by giant transposable elements. Trends Genet **21:** 551-552.

FESCHOTTE, C., X. ZHANG and S. WESSLER, 2002a Miniature inverted-repeat transposable elements and their relationship to established DNA transposons, pp. 1147–1158 in *Mobile DNA II*, edited by N. CRAIG, R. CRAIGIE, M. GELLERT and A. LAMBOWITZ. American Society for Microbiology Press, Washington, DC.

FESCHOTTE, C., X. ZHANG and S. R. WESSLER, 2002b Miniature inverted-repeat transposable elements and their relationships to established DNA transposons.

FEUCHTER-MURTHY, A. E., J. D. FREEMAN and D. L. MAGER, 1993 Splicing of a human endogenous retrovirus to a novel phospholipase A2 related gene. Nucleic Acids Res **21:** 135-143.

FEUCHTER, A. E., J. D. FREEMAN and D. L. MAGER, 1992 Strategy for detecting cellular transcripts promoted by human endogenous long terminal repeats: identification of a novel gene (CDC4L) with homology to yeast CDC4. Genomics **13:** 1237-1246.

FINNEGAN, D. J., 1989 Eukaryotic transposable elements and genome evolution. Trends Genet **5:** 103-107.

FINNEGAN, D. J., 1992 Transposable elements. Curr Opin Genet Dev **2:** 861-867.

FIRE, A., S. XU, M. K. MONTGOMERY, S. A. KOSTAS, S. E. DRIVER *et al.*, 1998 Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature **391:** 806-811.

FLAVELL, A. J., S. R. PEARCE and A. KUMAR, 1994 Plant transposable elements and the genome. Curr Opin Genet Dev **4:** 838-844.

FLAVELL, R. B., M. D. BENNETT, J. B. SMITH and D. B. SMITH, 1974 Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet **12:** 257-269.

FORNASARI, D., E. BATTAGLIOLI, A. FLORA, S. TERZANO and F. CLEMENTI, 1997 Structural and functional characterization of the human alpha3 nicotinic subunit gene promoter. Mol Pharmacol **51:** 250-261.

FRIESEN, P. D., W. C. RICE, D. W. MILLER and L. K. MILLER, 1986 Bidirectional transcription from a solo long terminal repeat of the retrotransposon TED: symmetrical RNA start sites. Mol Cell Biol **6:** 1599-1607.

GALE, M., JR., C. M. BLAKELY, D. A. HOPKINS, M. W. MELVILLE, M. WAMBACH *et al.*, 1998 Regulation of interferon-induced protein kinase PKR: modulation of P58IPK inhibitory function by a novel protein, P52rIPK. Mol Cell Biol **18:** 859-871.

GAO, X., and D. F. VOYTAS, 2005 A eukaryotic gene family related to retroelement integrases. Trends Genet **21:** 133-137.

GARDNER, M. J., N. HALL, E. FUNG, O. WHITE, M. BERRIMAN *et al.*, 2002 Genome sequence of the human malaria parasite Plasmodium falciparum. Nature **419:** 498-511.

GIBBS, R. A., G. M. WEINSTOCK, M. L. METZKER, D. M. MUZNY, E. J. SODERGREN *et al.*, 2004 Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428:** 493-521.

GLOCKNER, G., K. SZAFRANSKI, T. WINCKLER, T. DINGERMANN, M. A. QUAIL *et al.*, 2001 The complex repeats of Dictyostelium discoideum. Genome Res **11:** 585-594.

GOODCHILD, N. L., D. A. WILKINSON and D. L. MAGER, 1992 A human endogenous long terminal repeat provides a polyadenylation signal to a novel, alternatively spliced transcript in normal placenta. Gene **121:** 287-294.

GOTEA, V., and W. MAKALOWSKI, 2006 Do transposable elements really contribute to proteomes? Trends Genet **22:** 260-267.

GOULD, S. J., and E. S. VRBA, 1982 Exaptation: a missing term in the science of form. Paleobiology **8:** 4-15.

GRIFFITHS-JONES, S., R. J. GROCOCK, S. VAN DONGEN, A. BATEMAN and A. J. ENRIGHT, 2006 miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res **34:** D140-144.

GRUENBERG, B. H., A. SCHOENEMEYER, B. WEISS, L. TOSCHI, S. KUNZ *et al.*, 2001 A novel, soluble homologue of the human IL-10 receptor with preferential expression in placenta. Genes Immun **2:** 329-334.

HAMBOR, J. E., J. MENNONE, M. E. COON, J. H. HANKE and P. KAVATHAS, 1993 Identification and characterization of an Alu-containing, T-cell-specific enhancer located in the last intron of the human CD8 alpha gene. Mol Cell Biol **13:** 7056-7070.

HAMDI, H. K., H. NISHIO, J. TAVIS, R. ZIELINSKI and A. DUGAICZYK, 2000 Alu-mediated phylogenetic novelties in gene regulation and development. J Mol Biol **299:** 931-939.

HAMILTON, A., O. VOINNET, L. CHAPPELL and D. BAULCOMBE, 2002 Two classes of short interfering RNA in RNA silencing. Embo J **21:** 4671-4679.

HAMMER, S. E., S. STREHL and S. HAGEMANN, 2005 Homologs of Drosophila P transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human. Mol Biol Evol **22:** 833-844.

HAN, K., M. K. KONKEL, J. XING, H. WANG, J. LEE *et al.*, 2007 Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. Science **316:** 238-240.

HARENDZA, C. J., and L. F. JOHNSON, 1990 Polyadenylylation signal of the mouse thymidylate synthase gene was created by insertion of an L1 repetitive element downstream of the open reading frame. Proc Natl Acad Sci U S A **87:** 2531-2535.

HARTL, D. L., D. E. DYKHUIZEN, R. D. MILLER, L. GREEN and J. DE FRAMOND, 1983 Transposable element IS50 improves growth rate of E. coli cells without transposition. Cell **35:** 503-510.

HEWITT, S. M., G. C. FRAIZER and G. F. SAUNDERS, 1995 Transcriptional silencer of the Wilms' tumor gene WT1 contains an Alu repeat. J Biol Chem **270:** 17908-17912.

HICKEY, D. A., 1982 Selfish DNA: a sexually-transmitted nuclear parasite. Genetics **101:** 519-531.

HIOM, K., M. MELEK and M. GELLERT, 1998 DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell **94:** 463-470.

HOFACKER, I., W. FONTANA, P. STADLER, S. BONHOEFFER, M. TACKER *et al.*, 1994 Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125:** 167-188.

HOHMANN, S., 1993 Characterisation of PDC2, a gene necessary for high level expression of pyruvate decarboxylase structural genes in Saccharomyces cerevisiae. Mol Gen Genet **241:** 657-666.

HOLT, R. A., G. M. SUBRAMANIAN, A. HALPERN, G. G. SUTTON, R. CHARLAB *et al.*, 2002 The genome sequence of the malaria mosquito Anopheles gambiae. Science **298:** 129-149.

HUANG, J., Q. MORRIS and B. FREY, 2006 Detecting microRNA targets by linking sequence, microRNA and gene expression data, pp. 114-129 in *RECOMB 2006*, edited by A. APOSTOLICO, C. GUERRA, S. ISTRAIL, P. PEVZNER and M. WATERMAN. Springer-Verlag, Venice, Italy.

HUDSON, M. E., D. R. LISCH and P. H. QUAIL, 2003 The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant J **34:** 453-471.

HUMPHREY, G. W., E. W. ENGLANDER and B. H. HOWARD, 1996 Specific binding sites for a pol III transcriptional repressor and pol II transcription factor YY1 within the internucleosomal spacer region in primate Alu repetitive elements. Gene Expr **6:** 151-168.

HUTTENHOFER, A., and J. VOGEL, 2006 Experimental approaches to identify non-coding RNAs. Nucleic Acids Res **34:** 635-646.

HUTVAGNER, G., and P. D. ZAMORE, 2002a A microRNA in a multiple-turnover RNAi enzyme complex. Science **297:** 2056-2060.

HUTVAGNER, G., and P. D. ZAMORE, 2002b RNAi: nature abhors a double-strand. Curr Opin Genet Dev **12:** 225-232.

INTERNATIONAL CHICKEN GENOME SEQUENCING CONSORTIUM, 2004 Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432:** 695-716.

JABBARI, K., S. CRUVEILLER, O. CLAY, J. LE SAUX and G. BERNARDI, 2004 The new genes of rice: a closer look. Trends Plant Sci **9:** 281-285.

JASINSKA, A., and W. J. KRZYZOSIAK, 2004 Repetitive sequences that shape the human transcriptome. FEBS Lett **567:** 136-141.

JIANG, N., C. FESCHOTTE, X. ZHANG and S. R. WESSLER, 2004 Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr Opin Plant Biol **7:** 115-119.

JORDAN, I. K., I. B. ROGOZIN, G. V. GLAZKO and E. V. KOONIN, 2003 Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet **19:** 68-72.

JURKA, J., 2000 Repbase update: a database and an electronic journal of repetitive elements. Trends Genet **16:** 418-420.

JURKA, J., 2006 MER135: conserved mammalian repeat, probably derived from a non-autonomous DNA transposon. Repbase Rep. **6:** 388.

JURKA, J., and V. V. KAPITONOV, 1999 Sectorial mutagenesis by transposable elements. Genetica **107:** 239-248.

JURKA, J., V. V. KAPITONOV, O. KOHANY and M. V. JURKA, 2007 Repetitive Sequences in Complex Genomes: Structure and Evolution. Annu Rev Genomics Hum Genet.

JURKA, J., V. V. KAPITONOV, A. PAVLICEK, P. KLONOWSKI, O. KOHANY *et al.*, 2005 Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res **110:** 462-467.

KAMAL, M., X. XIE and E. S. LANDER, 2006 A large family of ancient repeat elements in the human genome is under strong selection. Proc Natl Acad Sci U S A **103:** 2740-2745.

KAPITONOV, V. V., and J. JURKA, 2004 Harbinger transposons and an ancient HARBI1 gene derived from a transposase. DNA Cell Biol **23:** 311-324.

KAPITONOV, V. V., and J. JURKA, 2005 RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol **3:** e181.

KAPITONOV, V. V., A. PAVLICEK and J. JURKA, 2004 Anthology of human repetitive DNA, pp. 251-305 in *Encyclopedia of molecular cell biology and molecular medicine*, edited by R. A. MEYERS. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

KAROLCHIK, D., R. BAERTSCH, M. DIEKHANS, T. S. FUREY, A. HINRICHS *et al.*, 2003 The UCSC Genome Browser Database. Nucleic Acids Res **31:** 51-54.

KAROLCHIK, D., A. S. HINRICHS, T. S. FUREY, K. M. ROSKIN, C. W. SUGNET *et al.*, 2004 The UCSC Table Browser data retrieval tool. Nucleic Acids Res **32:** D493-496.

KATO, N., K. SHIMOTOHNO, D. VANLEEUWEN and M. COHEN, 1990 Human proviral mRNAs down regulated in choriocarcinoma encode a zinc finger protein related to Kruppel. Mol Cell Biol **10:** 4401-4405.

KAZAKOV, V. I., and N. V. TOMILIN, 1996 Increased concentration of some transcription factor binding sites in human retroposons of the Alu family. Genetica **97:** 15-22.

KAZAZIAN, H. H., JR., 1998 Mobile elements and disease. Curr Opin Genet Dev **8:** 343-350.

KAZAZIAN, H. H., JR., 2004 Mobile elements: drivers of genome evolution. Science **303:** 1626-1632.

KENT, W. J., 2002 BLAT--the BLAST-like alignment tool. Genome Res **12:** 656-664.

KENT, W. J., R. BAERTSCH, A. HINRICHS, W. MILLER and D. HAUSSLER, 2003 Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A **100:** 11484-11489.

KENT, W. J., C. W. SUGNET, T. S. FUREY, K. M. ROSKIN, T. H. PRINGLE *et al.*, 2002 The human genome browser at UCSC. Genome Res **12:** 996-1006.

KETTING, R. F., T. H. HAVERKAMP, H. G. VAN LUENEN and R. H. PLASTERK, 1999 Mut-7 of C. elegans, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. Cell **99:** 133-141.

KIDO, S., N. SAKURAGI, M. P. BRONNER, R. SAYEGH, R. BERGER *et al.*, 1993 D21S418E identifies a cAMP-regulated gene located on chromosome 21q22.3 that is expressed in placental syncytiotrophoblast and choriocarcinoma cells. Genomics **17:** 256-259.

KIDWELL, M. G., 2002 Transposable elements and the evolution of genome size in eukaryotes. Genetica **115:** 49-63.

KIDWELL, M. G., and D. LISCH, 1997 Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A **94:** 7704-7711.

KIDWELL, M. G., and D. R. LISCH, 2001 Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution **55:** 1-24.

KIM, J. H., C. Y. YU, A. BAILEY, R. HARDISON and C. K. SHEN, 1989 Unique sequence organization and erythroid cell-specific nuclear factor-binding of mammalian theta 1 globin promoters. Nucleic Acids Res **17:** 5687-5700.

KIM, J. M., S. VANGURI, J. D. BOEKE, A. GABRIEL and D. F. VOYTAS, 1998 Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome Res **8:** 464-478.

KIPLING, D., and P. E. WARBURTON, 1997 Centromeres, CENP-B and Tigger too. Trends Genet **13:** 141-145.

KITAHARA, O., Y. FURUKAWA, T. TANAKA, C. KIHARA, K. ONO *et al.*, 2001 Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumor tissues and normal epithelia. Cancer Res **61:** 3544-3549.

KNIGHT, S. W., and B. L. BASS, 2001 A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. Science **293:** 2269-2271.

KOTENKO, S. V., L. S. IZOTOVA, O. V. MIROCHNITCHENKO, E. ESTEROVA, H. DICKENSHEETS *et al.*, 2001 Identification, cloning, and characterization of a novel soluble receptor that binds IL-22 and neutralizes its activity. J Immunol **166:** 7096-7103.

KRESS, M., Y. BARRA, J. G. SEIDMAN, G. KHOURY and G. JAY, 1984 Functional insertion of an Alu type 2 (B2 SINE) repetitive sequence in murine class I genes. Science **226:** 974-977.

KRIEGS, J. O., J. SCHMITZ, W. MAKALOWSKI and J. BROSIUS, 2005 Does the AD7c-NTP locus encode a protein? Biochim Biophys Acta **1727:** 1-4.

KRULL, M., J. BROSIUS and J. SCHMITZ, 2005 Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol **22:** 1702-1711.

KUMAR, A., and J. L. BENNETZEN, 1999 Plant retrotransposons. Annu Rev Genet **33:** 479-532.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform **5:** 150-163.

KWON, H. C., S. H. KIM, M. S. ROH, J. S. KIM, H. S. LEE *et al.*, 2004 Gene expression profiling in lymph node-positive and lymph node-negative colorectal cancer. Dis Colon Rectum **47:** 141-152.

LAGOS-QUINTANA, M., R. RAUHUT, W. LENDECKEL and T. TUSCHL, 2001 Identification of novel genes coding for small expressed RNAs. Science **294:** 853-858.

LAIMINS, L., M. HOLMGREN-KONIG and G. KHOURY, 1986 Transcriptional "silencer" element in rat repetitive sequences associated with the rat insulin 1 gene locus. Proc Natl Acad Sci U S A **83:** 3151-3155.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LANDRY, J. R., D. L. MAGER and B. T. WILHELM, 2003 Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet **19:** 640-648.

LANDRY, J. R., A. ROUHI, P. MEDSTRAND and D. L. MAGER, 2002 The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. Mol Biol Evol **19:** 1934-1942.

LARKIN, D. M., A. EVERTS-VAN DER WIND, M. REBEIZ, P. A. SCHWEITZER, S. BACHMAN *et al.*, 2003 A cattle-human comparative map built with cattle BAC-ends and human genome sequence. Genome Res **13:** 1966-1972.

LAU, N. C., L. P. LIM, E. G. WEINSTEIN and D. P. BARTEL, 2001 An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science **294:** 858-862.

LEE, R. C., and V. AMBROS, 2001 An extensive class of small RNAs in Caenorhabditis elegans. Science **294:** 862-864.

LEE, R. C., R. L. FEINBAUM and V. AMBROS, 1993 The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell **75:** 843-854.

LEE, S. H., M. OSHIGE, S. T. DURANT, K. K. RASILA, E. A. WILLIAMSON *et al.*, 2005 The SET domain protein Metnase mediates foreign DNA integration and links integration to nonhomologous end-joining repair. Proc Natl Acad Sci U S A **102:** 18075-18080.

LEE, Y., K. JEON, J. T. LEE, S. KIM and V. N. KIM, 2002 MicroRNA maturation: stepwise processing and subcellular localization. EMBO J **21:** 4663-4670.

LI, S. C., C. Y. PAN and W. C. LIN, 2006 Bioinformatic discovery of microRNA precursors from human ESTs and introns. BMC Genomics **7:** 164.

LINDBLAD-TOH, K., C. M. WADE, T. S. MIKKELSEN, E. K. KARLSSON, D. B. JAFFE *et al.*, 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature **438:** 803-819.

LINDBO, J. A., L. SILVA-ROSALES, W. M. PROEBSTING and W. G. DOUGHERTY, 1993 Induction of a Highly Specific Antiviral State in Transgenic Plants: Implications for Regulation of Gene Expression and Virus Resistance. Plant Cell **5:** 1749-1759.

LINDOW, M., and A. KROGH, 2005 Computational evidence for hundreds of non-conserved plant microRNAs. BMC Genomics **6:** 119.

LINGNER, J., T. R. HUGHES, A. SHEVCHENKO, M. MANN, V. LUNDBLAD *et al.*, 1997 Reverse transcriptase motifs in the catalytic subunit of telomerase. Science **276:** 561-567.

LIPPMAN, Z., A. V. GENDREL, M. BLACK, M. W. VAUGHN, N. DEDHIA *et al.*, 2004 Role of transposable elements in heterochromatin and epigenetic control. Nature **430:** 471-476.

LIPPMAN, Z., B. MAY, C. YORDAN, T. SINGER and R. MARTIENSSEN, 2003 Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. PLoS Biol **1:** E67.

LIU, A. Y., and B. A. ABRAHAM, 1991 Subtractive cloning of a hybrid human endogenous retrovirus and calbindin gene in the prostate cell line PC3. Cancer Res **51:** 4107-4110.

LIU, A. Y., and R. C. BRADNER, 1993 Elevated expression of the human mitochondrial hinge protein gene in cancer. Cancer Res **53:** 2460-2465.

LLAVE, C., K. D. KASSCHAU, M. A. RECTOR and J. C. CARRINGTON, 2002 Endogenous and silencing-associated small RNAs in plants. Plant Cell **14:** 1605-1619.

LLORENS, C., and I. MARIN, 2001 A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol **18:** 1597-1600.

LORENC, A., and W. MAKALOWSKI, 2003 Transposable elements and vertebrate protein diversity. Genetica **118:** 183-191.

LU, C., K. KULKARNI, F. F. SOURET, R. MUTHUVALLIAPPAN, S. S. TEJ *et al.*, 2006 MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. Genome Res **16:** 1276-1288.

LUKASHIN, A. V., and M. BORODOVSKY, 1998 GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res **26:** 1107-1115.

LYNCH, M., 2007 *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.

MAGER, D. L., 1989 Polyadenylation function and sequence variability of the long terminal repeats of the human endogenous retrovirus-like family RTVL-H. Virology **173:** 591-599.

MAICHELE, A. J., N. J. FARWELL and J. S. CHAMBERLAIN, 1993 A B2 repeat insertion generates alternate structures of the mouse muscle gamma-phosphorylase kinase gene. Genomics **16:** 139-149.

MAKALOWSKI, W., 1995 SINEs as a genomic scrap yard: an essay on genomic evolution, pp. 81-104 in *The impact of short interspersed elements (SINEs) on the host genome*, edited by R. J. MARAIA. R.G.Landes, Austin, Texas.

MAKALOWSKI, W., 2000 Genomic scrap yard: how genomes utilize all that junk. Gene **259:** 61-67.

MAKALOWSKI, W., 2003 Genomics. Not junk after all. Science **300:** 1246-1247.

MAKALOWSKI, W., G. A. MITCHELL and D. LABUDA, 1994 Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet **10:** 188-193.

MAKALOWSKI, W., and Y. TODA, 2007 Modulation of host genes by mammalian transposable elements. Genome Dynamics **3:** 163-175.

MALIK, H. S., and S. HENIKOFF, 2005 Positive selection of Iris, a retroviral envelope-derived host gene in Drosophila melanogaster. PLoS Genet **1:** e44.

MAO, L., T. C. WOOD, Y. YU, M. A. BUDIMAN, J. TOMKINS *et al.*, 2000 Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res **10:** 982-990.

MARINO-RAMIREZ, L., and I. K. JORDAN, 2006 Transposable element derived DNaseI-hypersensitive sites in the human genome. Biol Direct **1:** 20.

MARINO-RAMIREZ, L., K. C. LEWIS, D. LANDSMAN and I. K. JORDAN, 2005 Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogenet Genome Res **110:** 333-341.

MARKER, C., A. ZEMANN, T. TERHORST, M. KIEFMANN, J. P. KASTENMAYER *et al.*, 2002 Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant Arabidopsis thaliana. Curr Biol **12:** 2002-2013.

MARTINEZ, J., A. PATKANIOWSKA, H. URLAUB, R. LUHRMANN and T. TUSCHL, 2002 Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell **110:** 563-574.

MATSUMINE, H., M. A. HERBST, S. H. OU, J. D. WILSON and M. J. MCPHAUL, 1991 Aromatase mRNA in the extragonadal tissues of chickens with the henny-feathering trait is derived from a distinctive promoter structure that contains a segment of a retroviral long terminal repeat. Functional organization of the Sebright, Leghorn, and Campine aromatase genes. J Biol Chem **266:** 19900-19907.

MATTICK, J. S., and I. V. MAKUNIN, 2006 Non-coding RNA. Hum Mol Genet **15 Spec No 1:** R17-29.

MATZKE, M. A., and A. J. MATZKE, 2004 Planting the seeds of a new paradigm. PLoS Biol **2:** E133.

MATZKE, M. A., M. F. METTE and A. J. MATZKE, 2000 Transgene silencing by the host genome defense: implications for the evolution of epigenetic control mechanisms in plants and vertebrates. Plant Mol Biol **43:** 401-415.

MCCLINTOCK, B., 1948 Mutable loci in maize. Carnegie Inst Wash Year Book **47:** 155-169.

MCCLINTOCK, B., 1984 The significance of responses of the genome to challenge. Science **226:** 792-801.

MCDONALD, J. F., 1993 Evolution and consequences of transposable elements. Curr Opin Genet Dev **3:** 855-864.

MCDONALD, J. F., 1995 Transposable elements: possible catalysts of organismic evolution. Trends Ecol Evol **10:** 123-126.

MCDONALD, J. F., 1999 Genomic imprinting as a coopted evolutionary character. Trends Ecol Evol **14:** 359.

MCDONALD, J. F., M. A. MATZKE and A. J. MATZKE, 2005 Host defenses to transposable elements and the evolution of genomic imprinting. Cytogenet Genome Res **110:** 242-249.

MCHAFFIE, G. S., and S. H. RALSTON, 1995 Origin of a negative calcium response element in an ALU-repeat: implications for regulation of gene expression by extracellular calcium. Bone **17:** 11-14.

MEDSTRAND, P., J. R. LANDRY and D. L. MAGER, 2001 Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem **276:** 1896-1903.

MEDSTRAND, P., L. N. VAN DE LAGEMAAT, C. A. DUNN, J. R. LANDRY, D. SVENBACK *et al.*, 2005 Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res **110:** 342-352.

METTE, M. F., J. VAN DER WINDEN, M. MATZKE and A. J. MATZKE, 2002 Short RNAs can identify new candidate transposable element families in Arabidopsis. Plant Physiol **130:** 6-9.

MEYERS, B. C., D. K. LEE, T. H. VU, S. S. TEJ, S. B. EDBERG *et al.*, 2004 Arabidopsis MPSS. An online resource for quantitative expression analysis. Plant Physiol **135:** 801-813.

MI, S., X. LEE, X. LI, G. M. VELDMAN, H. FINNERTY *et al.*, 2000 Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature **403:** 785-789.

MIGHELL, A. J., A. F. MARKHAM and P. A. ROBINSON, 1997 Alu sequences. FEBS Lett **417:** 1-5.

MIKKELSEN, T., W. LADEANA, E. E. EICHLER, M. C. ZODY, D. B. JAFFE *et al.*, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437:** 69-87.

MILLER, W. J., S. HAGEMANN, E. REITER and W. PINSKER, 1992 P-element homologous sequences are tandemly repeated in the genome of Drosophila guanche. Proc Natl Acad Sci U S A **89:** 4018-4022.

MORGAN, G. T., 1995 Identification in the human genome of mobile elements spread by DNA-mediated transposition. J Mol Biol **254:** 1-5.

MOURELATOS, Z., J. DOSTIE, S. PAUSHKIN, A. SHARMA, B. CHARROUX *et al.*, 2002 miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. Genes Dev **16:** 720-728.

MUEHLBAUER, G. J., B. S. BHAU, N. H. SYED, S. HEINEN, S. CHO *et al.*, 2006 A hAT superfamily transposase recruited by the cereal grass genome. Mol Genet Genomics **275:** 553-563.

MUIR, A., A. LEVER and A. MOFFETT, 2004 Expression and functions of human endogenous retroviruses in the placenta: an update. Placenta **25 Suppl A:** S16-25.

MURNANE, J. P., and J. F. MORALES, 1995 Use of a mammalian interspersed repetitive (MIR) element in the coding and processing sequences of mammalian genes. Nucleic Acids Res **23:** 2837-2839.

NAGASAKI, K., C. SCHEM, C. VON KAISENBERG, M. BIALLEK, F. ROSEL *et al.*, 2003 Leucine-zipper protein, LDOC1, inhibits NF-kappaB activation and sensitizes pancreatic cancer cells to apoptosis. Int J Cancer **105:** 454-458.

NAKAMURA, T. M., G. B. MORIN, K. B. CHAPMAN, S. L. WEINRICH, W. H. ANDREWS *et al.*, 1997 Telomerase catalytic subunit homologs from fission yeast and human. Science **277:** 955-959.

NAM, J. W., K. R. SHIN, J. HAN, Y. LEE, V. N. KIM *et al.*, 2005 Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res **33:** 3570-3581.

NAPOLI, C., C. LEMIEUX and R. JORGENSEN, 1990 Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. Plant Cell **2:** 279-289.

NEI, M., and S. KUMAR, 2000 *Molecular Evolution and Phylogenetics*. Oxford, New York.

NEKRUTENKO, A., and W. H. LI, 2001 Transposable elements are found in a large number of human protein-coding genes. Trends Genet **17:** 619-621.

NEZNANOV, N. S., and R. G. OSHIMA, 1993 cis regulation of the keratin 18 gene in transgenic mice. Mol Cell Biol **13:** 1815-1823.

NISHIHARA, H., A. F. SMIT and N. OKADA, 2006 Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res **16:** 864-874.

NOBUTA, K., R. C. VENU, C. LU, A. BELO, K. VEMARAJU *et al.*, 2007 An expression atlas of rice mRNAs and small RNAs. Nat Biotechnol **25:** 473-477.

NORRIS, J., D. FAN, C. ALEMAN, J. R. MARKS, P. A. FUTREAL *et al.*, 1995 Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. J Biol Chem **270:** 22777-22782.

NOTTERMAN, D. A., U. ALON, A. J. SIERK and A. J. LEVINE, 2001 Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res **61:** 3124-3130.

OEI, S. L., J. GRIESENBECK, M. SCHWEIGER, V. BABICH, A. KROPOTOV *et al.*, 1997 Interaction of the transcription factor YY1 with human poly(ADP-ribosyl) transferase. Biochem Biophys Res Commun **240:** 108-111.

OHNO, S., 1972 So much "junk" DNA in our genome. Brookhaven Symp. Biol. **23:** 366-370.

OLIVIERO, S., and P. MONACI, 1988 RNA polymerase III promoter elements enhance transcription of RNA polymerase II genes. Nucleic Acids Res **16:** 1285-1293.

ONO, R., K. NAKAMURA, K. INOUE, M. NARUSE, T. USAMI *et al.*, 2006 Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet **38:** 101-106.

OOSUMI, T., W. R. BELKNAP and B. GARLICK, 1995 Mariner transposons in humans. Nature **378:** 672.

ORGEL, L. E., and F. H. CRICK, 1980 Selfish DNA: the ultimate parasite. Nature **284:** 604-607.

OUYANG, S., W. ZHU, J. HAMILTON, H. LIN, M. CAMPBELL *et al.*, 2007 The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res **35:** D883-887.

PASQUINELLI, A. E., B. J. REINHART, F. SLACK, M. Q. MARTINDALE, M. I. KURODA *et al.*, 2000 Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature **408:** 86-89.

PAULSON, K. E., A. G. MATERA, N. DEKA and C. W. SCHMID, 1987 Transcription of a human transposon-like sequence is usually directed by other promoters. Nucleic Acids Res **15:** 5199-5215.

PAVLICEK, A., O. CLAY and G. BERNARDI, 2002 Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? FEBS Lett **523:** 252-253.

PEDERSEN, J. S., G. BEJERANO, A. SIEPEL, K. ROSENBLOOM, K. LINDBLAD-TOH *et al.*, 2006 Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol **2:** e33.

PIEDRAFITA, F. J., R. B. MOLANDER, G. VANSANT, E. A. ORLOVA, M. PFAHL *et al.*, 1996 An Alu element in the myeloperoxidase promoter contains a composite SP1-thyroid hormone-retinoic acid response element. J Biol Chem **271:** 14412-14420.

PINSKER, W., E. HARING, S. HAGEMANN and W. J. MILLER, 2001 The evolutionary life history of P transposons: from horizontal invaders to domesticated neogenes. Chromosoma **110:** 148-158.

PIRIYAPONGSA, J., and I. K. JORDAN, 2007 A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS ONE **2:** e203.

PIRIYAPONGSA, J., L. MARINO-RAMIREZ and I. K. JORDAN, 2007a Origin and evolution of human microRNAs from transposable elements. Genetics **176:** 1323-1337.

PIRIYAPONGSA, J., N. POLAVARAPU, M. BORODOVSKY and J. MCDONALD, 2007b Exonization of the LTR transposable elements in human genome. BMC Genomics **8:** 291.

PLANT, K. E., S. J. ROUTLEDGE and N. J. PROUDFOOT, 2001 Intergenic transcription in the human beta-globin gene cluster. Mol Cell Biol **21:** 6507-6514.

PLASTERK, R. H., 2002 RNA silencing: the genome's immune system. Science **296:** 1263-1265.

POLLARD, K. S., S. R. SALAMA, B. KING, A. D. KERN, T. DRESZER *et al.*, 2006a Forces shaping the fastest evolving regions in the human genome. PLoS Genet **2:** e168.

POLLARD, K. S., S. R. SALAMA, N. LAMBERT, M. A. LAMBOT, S. COPPENS *et al.*, 2006b An RNA gene expressed during cortical development evolved rapidly in humans. Nature **443:** 167-172.

POULTER, R., and M. BUTLER, 1998 A retrotransposon family from the pufferfish (fugu) Fugu rubripes. Gene **215:** 241-249.

PRABHAKAR, S., J. P. NOONAN, S. PAABO and E. M. RUBIN, 2006 Accelerated evolution of conserved noncoding sequences in humans. Science **314:** 786.

QUESNEVILLE, H., D. NOUAUD and D. ANXOLABEHERE, 2005 Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element. Mol Biol Evol **22:** 741-746.

RAMAKRISHNAN, C., and D. M. ROBINS, 1997 Steroid hormone responsiveness of a family of closely related mouse proviral elements. Mamm Genome **8:** 811-817.

RATCLIFF, F., B. D. HARRISON and D. C. BAULCOMBE, 1997 A similarity between viral defense and gene silencing in plants **275**.

REINHART, B. J., F. J. SLACK, M. BASSON, A. E. PASQUINELLI, J. C. BETTINGER *et al.*, 2000 The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature **403:** 901-906.

REISS, D., D. NOUAUD, S. RONSSERAY and D. ANXOLABEHERE, 2005 Domesticated P elements in the Drosophila montium species subgroup have a new function related to a DNA binding property. J Mol Evol **61:** 470-480.

RHOADES, M. W., B. J. REINHART, L. P. LIM, C. B. BURGE, B. BARTEL *et al.*, 2002 Prediction of plant microRNA targets. Cell **110:** 513-520.

ROBERTSON, H. M., 2002 Evolution of DNA transposons in eukaryotes, pp. 1093-1110 in *Mobile DNA II*, edited by N. CRAIG, R. CRAIGIE, M. GELLERT and A. LAMBOWITZ. ASM Press, Washington.

ROBERTSON, H. M., and W. R. ENGELS, 1989 Modified P elements that mimic the P cytotype in Drosophila melanogaster. Genetics **123:** 815-824.

ROBERTSON, H. M., and K. L. ZUMPANO, 1997 Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome. Gene **205:** 203-217.

ROBINS, D. M., and L. C. SAMUELSON, 1992 Retrotransposons and the evolution of mammalian gene expression. Genetica **86:** 191-201.

ROSE, M. R., and W. F. DOOLITTLE, 1983 Molecular Biological Mechanisms of Speciation. Science **220:** 157-162.

ROTHKOPF, G. S., C. A. TELAKOWSKI-HOPKINS, R. L. STOTISH and C. B. PICKETT, 1986 Multiplicity of glutathione S-transferase genes in the rat and association with a type 2 Alu repetitive element. Biochemistry **25:** 993-1002.

ROUSSIGNE, M., S. KOSSIDA, A. C. LAVIGNE, T. CLOUAIRE, V. ECOCHARD *et al.*, 2003 The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. Trends Biochem Sci **28:** 66-69.

ROZEN, S., and H. J. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers, pp. 365-386 in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, edited by S. MISENER and S. A. KRAWETZ. Humana Press, Totowa.

RUBY, J. G., C. H. JAN and D. P. BARTEL, 2007 Intronic microRNA precursors that bypass Drosha processing. Nature **448:** 83-86.

RYSKOV, A. P., P. L. IVANOV, D. A. KRAMEROV and G. P. GEORGIEV, 1984 [Universal orientation and 3'-terminal localization of repeated sequences in the B2 family of mRNA]. Mol Biol (Mosk) **18:** 92-103.

SAEGUSA, Y., M. SATO, I. GALLI, T. NAKAGAWA, N. ONO *et al.*, 1993 Stimulation of SV40 DNA replication and transcription by Alu family sequence. Biochim Biophys Acta **1172:** 274-282.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4:** 406-425.

SAKSELA, K., and D. BALTIMORE, 1993 Negative regulation of immunoglobulin kappa light-chain gene transcription by a short sequence homologous to the murine B1 repetitive element. Mol Cell Biol **13:** 3698-3705.

SAMUELSON, L. C., K. WIEBAUER, C. M. SNOW and M. H. MEISLER, 1990 Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. Mol Cell Biol **10:** 2513-2520.

SANMIGUEL, P., A. TIKHONOV, Y. K. JIN, N. MOTCHOULSKAIA, D. ZAKHAROV *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765-768.

SARKAR, A., C. SIM, Y. S. HONG, J. R. HOGAN, M. J. FRASER *et al.*, 2003 Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. Mol Genet Genomics **270:** 173-180.

SAUTER, M., S. SCHOMMER, E. KREMMER, K. REMBERGER, G. DOLKEN *et al.*, 1995 Human endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. J Virol **69:** 414-421.

SCHATZ, D. G., M. A. OETTINGER and D. BALTIMORE, 1989 The V(D)J recombination activating gene, RAG-1. Cell **59:** 1035-1048.

SCHULLER, M., D. JENNE and R. VOLTZ, 2005 The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. J Neuroimmunol **169:** 172-176.

SCHULTE, A. M., S. LAI, A. KURTZ, F. CZUBAYKO, A. T. RIEGEL *et al.*, 1996 Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. Proc Natl Acad Sci U S A **93:** 14759-14764.

SCHWARTZ, A., D. C. CHAN, L. G. BROWN, R. ALAGAPPAN, D. PETTAY *et al.*, 1998 Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Hum Mol Genet **7:** 1-11.

SCHWARZ, D. S., G. HUTVAGNER, B. HALEY and P. D. ZAMORE, 2002 Evidence that siRNAs function as guides, not primers, in the Drosophila and human RNAi pathways. Mol Cell **10:** 537-548.

SCHWEIGER, M., S. L. OEI, H. HERZOG, C. MENARDI, R. SCHNEIDER *et al.*, 1995 Regulation of the human poly(ADP-ribosyl) transferase promoter via alternative DNA racket structures. Biochimie **77:** 480-485.

SEITZ, H., N. YOUNGSON, S. P. LIN, S. DALBERT, M. PAULSEN *et al.*, 2003 Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. Nat Genet **34:** 261-262.

SELL, C., C. D. CHANG, J. KONIECKI, H. M. CHEN and R. BASERGA, 1992 A cryptopromoter is activated in the proliferating cell nuclear antigen gene of growth arrested cells. J Cell Physiol **152:** 177-184.

SEMIN, B. V., and V. IL'IN IU, 2005 [Diversity of LTR retrotransposons and their role in genome reorganization]. Genetika **41:** 542-548.

SHIH, W., R. CHETTY and M. S. TSAO, 2005 Expression profiling by microarrays in colorectal cancer (Review). Oncol Rep **13:** 517-524.

SIEPEL, A., G. BEJERANO, J. S. PEDERSEN, A. S. HINRICHS, M. HOU *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res **15:** 1034-1050.

SIJEN, T., and R. H. PLASTERK, 2003 Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. Nature **426:** 310-314.

SILVA, J. C., S. A. SHABALINA, D. G. HARRIS, J. L. SPOUGE and A. S. KONDRASHOVI, 2003 Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. Genet Res **82:** 1-18.

SINGER, S. S., D. N. MANNEL, T. HEHLGANS, J. BROSIUS and J. SCHMITZ, 2004 From "junk" to gene: curriculum vitae of a primate receptor isoform gene. J Mol Biol **341:** 883-886.

SLOTKIN, R. K., M. FREELING and D. LISCH, 2005 Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. Nat Genet **37:** 641-644.

SMALHEISER, N. R., and V. I. TORVIK, 2005 Mammalian microRNAs derived from genomic repeats. Trends Genet **21:** 322-326.

SMALHEISER, N. R., and V. I. TORVIK, 2006 Alu elements within human mRNAs are probable microRNA targets. Trends Genet **22:** 532-536.

SMIT, A., R. HUBLEY and P. GREEN, 1996-2004 RepeatMasker Open-3.0.

SMIT, A. F., 1993 Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res **21:** 1863-1872.

SMIT, A. F., 1999 Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev **9:** 657-663.

SMIT, A. F., and A. D. RIGGS, 1996 Tiggers and DNA transposon fossils in the human genome. Proc Natl Acad Sci U S A **93:** 1443-1448.

SMITH, T. F., and M. S. WATERMAN, 1981 Identification of common molecular subsequences. J Mol Biol **147:** 195-197.

SONNHAMMER, E. L., S. R. EDDY and R. DURBIN, 1997 Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins **28:** 405-420.

SOOD, P., A. KREK, M. ZAVOLAN, G. MACINO and N. RAJEWSKY, 2006 Cell-type-specific signatures of microRNAs on target mRNA expression. Proc Natl Acad Sci U S A **103:** 2746-2751.

SOREK, R., G. AST and D. GRAUR, 2002 Alu-containing exons are alternatively spliced. Genome Res **12:** 1060-1067.

STARK, A., J. BRENNECKE, N. BUSHATI, R. B. RUSSELL and S. M. COHEN, 2005 Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell **123:** 1133-1146.

STAVENHAGEN, J. B., and D. M. ROBINS, 1988 An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein gene. Cell **55:** 247-254.

STOLTZFUS, A., 2006 Mutationism and the dual causation of evolutionary change. Evol Dev **8:** 304-317.

STORZ, G., 2002 An expanding universe of noncoding RNAs. Science **296:** 1260-1263.

STURN, A., J. QUACKENBUSH and Z. TRAJANOSKI, 2002 Genesis: cluster analysis of microarray data. Bioinformatics **18:** 207-208.

SU, A. I., T. WILTSHIRE, S. BATALOV, H. LAPP, K. A. CHING *et al.*, 2004 A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A **101:** 6062-6067.

TABARA, H., M. SARKISSIAN, W. G. KELLY, J. FLEENOR, A. GRISHOK *et al.*, 1999 The rde-1 gene, RNA interference, and transposon silencing in C. elegans. Cell **99:** 123-132.

TAKEMASA, I., H. HIGUCHI, H. YAMAMOTO, M. SEKIMOTO, N. TOMITA *et al.*, 2001 Construction of preferential cDNA microarray specialized for human colorectal carcinoma: molecular sketch of colorectal cancer. Biochem Biophys Res Commun **285:** 1244-1249.

TAN, K. O., K. M. TAN, S. L. CHAN, K. S. YEE, M. BEVORT *et al.*, 2001 MAP-1, a novel proapoptotic protein containing a BH3-like motif that associates with Bax through its Bcl-2 homology domains. J Biol Chem **276:** 2802-2807.

TANG, G., B. J. REINHART, D. P. BARTEL and P. D. ZAMORE, 2003 A biochemical framework for RNA silencing in plants. Genes Dev **17:** 49-63.

THE ARABIDOPSIS GENOME INITIATIVE, 2000 Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796-815.

THE C. ELEGANS SEQUENCING CONSORTIUM, 1998 Genome sequence of the nematode C. elegans: a platform for investigating biology. Science **282:** 2012-2018.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-4680.

THOREY, I. S., G. CECENA, W. REYNOLDS and R. G. OSHIMA, 1993 Alu sequence involvement in transcriptional insulation of the keratin 18 gene in transgenic mice. Mol Cell Biol **13:** 6742-6751.

TING, C. N., M. P. ROSENBERG, C. M. SNOW, L. C. SAMUELSON and M. H. MEISLER, 1992 Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev **6:** 1457-1465.

TOMILIN, N. V., 1999 Control of genes by mammalian retroposons. Int Rev Cytol **186:** 1-48.

TOMILIN, N. V., S. M. IGUCHI-ARIGA and H. ARIGA, 1990 Transcription and replication silencer element is present within conserved region of human Alu repeats interacting with nuclear protein. FEBS Lett **263:** 69-72.

TONEGAWA, S., 1983 Somatic generation of antibody diversity. Nature **302:** 575-581.

TORARINSSON, E., M. SAWERA, J. H. HAVGAARD, M. FREDHOLM and J. GORODKIN, 2006 Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Res **16:** 885-889.

TOTH, M., J. GRIMSBY, G. BUZSAKI and G. P. DONOVAN, 1995 Epileptic seizures caused by inactivation of a novel gene, jerky, related to centromere binding protein-B in transgenic mice. Nat Genet **11:** 71-75.

TRELOGAN, S. A., and S. L. MARTIN, 1995 Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. Proc Natl Acad Sci U S A **92:** 1520-1524.

TURCOTTE, K., S. SRINIVASAN and T. BUREAU, 2001 Survey of transposable elements from rice genomic sequences. Plant J **25:** 169-179.

VAN BLOKLAND, R., N. VAN DER GEEST, J. N. M. MOL and J. M. KOOTER, 1994 Transgene-mediated suppression of chalcone synthase expression in Petunia hybrida results from an increase in RNA turnover. Plant J **6:** 861-877.

VAN DE LAGEMAAT, L. N., J. R. LANDRY, D. L. MAGER and P. MEDSTRAND, 2003 Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet **19:** 530-536.

VAN DER KROL, A. R., L. A. MUR, M. BELD, J. N. MOL and A. R. STUITJE, 1990 Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. Plant Cell **2:** 291-299.

VANSANT, G., and W. F. REYNOLDS, 1995 The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. Proc Natl Acad Sci U S A **92:** 8229-8233.

VASTENHOUW, N. L., and R. H. PLASTERK, 2004 RNAi protects the Caenorhabditis elegans germline against transposition. Trends Genet **20:** 314-319.

VICIENT, C. M., A. SUONIEMI, K. ANAMTHAWAT-JONSSON, J. TANSKANEN, A. BEHARAV *et al.*, 1999 Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum. Plant Cell **11:** 1769-1784.

VIEIRA, C., D. LEPETIT, S. DUMONT and C. BIEMONT, 1999 Wake up of transposable elements following Drosophila simulans worldwide colonization. Mol Biol Evol **16:** 1251-1255.

VOLFF, J. N., 2006 Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays **28:** 913-922.

VOLFF, J. N., and J. BROSIUS, 2007 Modern genomes with retro-look: retrotransposed elements, retroposition and the origin of new genes. Genome Dynamics **3:** 175-190.

WANG-JOHANNING, F., A. R. FROST, B. JIAN, L. EPP, D. W. LU *et al.*, 2003 Quantitation of HERV-K env gene expression and splicing in human breast cancer. Oncogene **22:** 1528-1535.

WASHIETL, S., I. L. HOFACKER, M. LUKASSER, A. HUTTENHOFER and P. F. STADLER, 2005a Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol **23:** 1383-1390.

WASHIETL, S., I. L. HOFACKER and P. F. STADLER, 2005b Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A **102:** 2454-2459.

WASSARMAN, K. M., F. REPOILA, C. ROSENOW, G. STORZ and S. GOTTESMAN, 2001 Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev **15:** 1637-1651.

WATERSTON, R. H., K. LINDBLAD-TOH, E. BIRNEY, J. ROGERS, J. F. ABRIL *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

WEISS, B., K. WOLK, B. H. GRUNBERG, H. D. VOLK, W. STERRY *et al.*, 2004 Cloning of murine IL-22 receptor alpha 2 and comparison with its human counterpart. Genes Immun **5:** 330-336.

WILLS, N. M., B. MOORE, A. HAMMER, R. F. GESTELAND and J. F. ATKINS, 2006 A functional -1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. J Biol Chem **281:** 7082-7088.

WILSON, C., P. GOETTING-MINESKY and A. NEKRUTENKO, 2006 mNSC1 shows no evidence of protein-coding capacity. Gene **370:** 83-85.

WU, J., G. J. GRINDLAY, P. BUSHEL, L. MENDELSOHN and M. ALLAN, 1990 Negative regulation of the human epsilon-globin gene by transcriptional interference: role of an Alu repetitive element. Mol Cell Biol **10:** 1209-1216.

XIE, X., M. KAMAL and E. S. LANDER, 2006 A family of conserved noncoding elements derived from an ancient transposable element. Proc Natl Acad Sci U S A **103:** 11659-11664.

XU, W., S. R. PRESNELL, J. PARRISH-NOVAK, W. KINDSVOGEL, S. JASPERS *et al.*, 2001 A soluble class II cytokine receptor, IL-22RA2, is a naturally occurring IL-22 antagonist. Proc Natl Acad Sci U S A **98:** 9511-9516.

YANG, Z., D. BOFFELLI, N. BOONMARK, K. SCHWARTZ and R. LAWN, 1998 Apolipoprotein(a) gene enhancer resides within a LINE element. J Biol Chem **273:** 891-897.

YEKTA, S., I. H. SHIH and D. P. BARTEL, 2004 MicroRNA-directed cleavage of HOXB8 mRNA. Science **304:** 594-596.

YODER, J. A., C. P. WALSH and T. H. BESTOR, 1997 Cytosine methylation and the ecology of intragenomic parasites. Trends Genet **13:** 335-340.

YOUNGSON, N. A., S. KOCIALKOWSKI, N. PEEL and A. C. FERGUSON-SMITH, 2005 A small family of sushi-class retrotransposon-derived genes in mammals and their relation to genomic imprinting. J Mol Evol **61:** 481-490.

YULUG, I. G., A. YULUG and E. M. FISHER, 1995 The frequency and position of Alu repeats in cDNAs, as determined by database searching. Genomics **27:** 544-548.

ZAISS, D. M., and P. M. KLOETZEL, 1999 A second gene encoding the mouse proteasome activator PA28beta subunit is part of a LINE1 element and is driven by a LINE1 promoter. J Mol Biol **287:** 829-835.

ZDOBNOV, E. M., M. CAMPILLOS, E. D. HARRINGTON, D. TORRENTS and P. BORK, 2005 Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. Nucleic Acids Res **33:** 946-954.

ZENG, Y., R. YI and B. R. CULLEN, 2003 MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. Proc Natl Acad Sci U S A **100:** 9779-9784.

ZENG, Z., H. KYAW, K. R. GAKENHEIMER, M. AUGUSTUS, P. FAN *et al.*, 1997 Cloning, mapping, and tissue distribution of a human homologue of the mouse jerky gene product. Biochem Biophys Res Commun **236:** 389-395.

ZHANG, B., D. SCHMOYER, S. KIROV and J. SNODDY, 2004 GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics **5:** 16.

ZHANG, Q., J. ARBUCKLE and S. R. WESSLER, 2000 Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize. Proc Natl Acad Sci U S A **97:** 1160-1165.

ZHANG, Z., and M. GERSTEIN, 2003 Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol **2:** 11.

ZHENG, J. H., S. NATSUUME-SAKAI, M. TAKAHASHI and M. NONAKA, 1992 Insertion of the B2 sequence into intron 13 is the only defect of the H-2k C4 gene which causes low C4 production. Nucleic Acids Res **20:** 4975-4979.

ZHOU, L., R. MITRA, P. W. ATKINSON, A. B. HICKMAN, F. DYDA *et al.*, 2004 Transposition of hAT elements links transposable elements and V(D)J recombination. Nature **432:** 995-1001.

ZHOU, Y. H., J. B. ZHENG, X. GU, G. F. SAUNDERS and W. K. YUNG, 2002 Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. Genome Res **12:** 1716-1722.

ZILBERMAN, D., X. CAO and S. E. JACOBSEN, 2003 ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. Science **299:** 716-719.

ZUKER, M., 2003 Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res **31:** 3406-3415.

VITA

JITTIMA PIRIYAPONGSA

Jittima Piriyapongsa was born in 1978 in Bangkok, Thailand. She received her Bachelor's Degree in Pharmacy from Chulalongkorn University, Thailand, in 2000. From 2000 to 2002, she worked as a researcher at the Chemistry Section of the Research and Development Institute, a part of the Government Pharmaceutical Organization, a state enterprise under Ministry of Public Health and the largest pharmaceutical manufacturer in Thailand. She was responsible for analyzing and developing the analytical methods for new pharmaceutical as well as phytopharmaceutical products using various equipments and techniques. In the end of 2002, she succeeded in being granted a scholarship from the Royal Thai Government to pursue her study in the United States, leading to PhD degree in the field of Bioinformatics. In August 2003, she decided to join the School of Biology at Georgia Institute of Technology. Under the conditions of the scholarship, she will return to serve as a researcher at The National Center for Genetic Engineering and Biotechnology (BIOTEC), one of the centers of The National Science and Technology Development Agency, Ministry of Science, Technology and Environment, Thailand.